Toward the Improved Simulation of Microscale Gas Flow

by

Matthew James McNenly

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Aerospace Engineering)
in The University of Michigan
2007

Doctoral Committee:

Professor Iain D. Boyd, Co-Chair
Professor Bram van Leer, Co-Chair
Associate Professor Luis P. Bernal
Assistant Professor Andrew J. Christlieb

To Anastacia

ACKNOWLEDGEMENTS

First and foremost, I must thank my thesis advisor Professor Iain Boyd. In addition to my gratitude, I probably owe him an apology for the rather winding path I traveled for my research, which must have appeared as a never-ending series of detours into Number Theory. I am thankful for the focus that his oversight provided me (along with the occasional mild threat), which helped me immensely to finish my thesis. Throughout his time as my advisor, Prof. Boyd remained eminently approachable for research discussions (even while on sabbatical), and served as a valuable source for new insight and guidance into my research problems. I am also very appreciative of his support and encouragement when it came to the other aspects of graduate school; specifically, applying for fellowships, presenting research at conferences, and submitting journal manuscripts.

For their time and consideration while serving on my thesis committee, I would like to thank: Professor Bram van Leer, Professor Andrew Christlieb and Professor Luis Bernal. Along with their helpful suggestions and criticism for my thesis, I am fortunate to have had many interesting discussions with them throughout my time as a graduate student, on a myriad of topics in gas dynamics and scientific computing. I must also reluctantly thank Prof. van Leer for his near impossible CFD II project, "Auroral Heating of Jupiter's Atmosphere." While the test conditions ultimately were deemed unstable, I am forced to admit at this time that the experience did

make me a better designer and user of numerical simulations.

I am grateful to the United States Department of Energy (DOE) for funding the majority of my graduate studies through the Computational Science Graduate Fellowship (CSGF) administered by the Krell Institute. I owe a tremendous debt to the CSGF program for my current achievement and for any future opportunities that I may enjoy. In my appreciation, I would like to thank Dr. James Corones, Barbara Helland, Shellie Hosch, Lucille Kilmer, Rachel Huisman and the rest of the Krell Institute for their skillful management of the CSGF program. I must also thank Pam Derry, who was equally deft coordinating the fellowship with the university.

The highlight of the CSGF program for me was clearly the summer practicum, which I spent at Sandia National Laboratories. I must thank Dr. Wahid Hermina for welcoming me to his talented research group (Dept. 9113) and for the support he provided during the practicum. At Sandia, I was impressed by the quality and breadth of the research being performed, which was due, in no small part, to the efforts of my mentor Dr. Michael Gallis. The research project he and I began ultimately led to my first published journal paper, and his contacts within the DOE community and his continuing interest in my career landed me my first job as a newly minted Ph. D. His time and energy spent helping me is so well appreciated that I fear a mere thanks does not adequately recognize his contribution.

I must thank all the graduate students that have helped me during my long tenure on the second floor of the FXB. My officemates and roommates deserve my greatest applause for putting up with me for so many hours in such a confined space: Dr. Jerold Emhoff II, Dr. Jesse Hoagg, Dr. Suhail Akhtar, Christian Morrison, Dr. Allen Victor, Anish Benjamin, and John Yim. I am grateful for the countless hours spent discussing our research, and for the new ideas and revelations these conver-

sations sparked. Similarly, I need to thank all the current and former members of the Nonequilibrium Gas and Plasma Dynamics (NGPD) Group for their comments, suggestions and criticism of my research; in particular, I would like to commend: Dave Berger, Jeremy Boerner, Dr. Jon Burt, Dr. Chunpei Cai, Yungjun Choi, Dr. Justin Koo, Dr. Michael Martin, Jose Padilla, Tom Schwartzentruber, Dr. Quanhua Sun, Dr. Anton VanderWyst, and Dr. Wen-Lan Wang. Also important are the extremely dedicated (or possibly nocturnal) grad students that I spent many nights studying alongside in the FXB: Dr. Stephen Broschart, Nalin Chaturvedi, Ashwani Padthe, Dr. Leonel Rios-Reyes, and Yoshifumi Suzuki. I am grateful for both their company and advice; however, I am probably most appreciative of their lax lending policies when it came time to buy some much needed Mt. Dew from the vending machine.

The staff of the Aerospace Department deserves special recognition. Throughout her tenure as the Graduate Student Services Coordinator (GSSC), Margaret Fillion devotedly shepherded her flock of often hapless students (especially in my case) through the countless forms needed by the university to remain paid, insured, and in this country. In addition to the patience and kindness that she always extended to me and rest of the graduate students, I must thank her for coming out of retirement to bake for my oral defense. I am also in the debt of the current GSSC, Denise Phelps, who managed to keep all my defense materials in Rackham's good graces, in spite of my paperwork management-style, or lack thereof. I would like to thank Debbie Laird, Diana McVey, Jenna McVey, Dorris Micou and Suzanne Smith as well for the many times they have cheerfully offered me assistance.

In my first semester as a graduate student, I would not have been able to survive my GSI assignment if it were not for the help of Dave McClean and Tom Griffin.

They gave me a crash course on equipment usage before every student lab, and were available on a moments notice to replace anything that broke in the middle of class. Without their support, I have no doubt that the undergrads would have revolted against me after the second lab assignment. I would also like to thank Chris Chartier and Eric Kirk for all their efforts maintaining the smooth operation of the FXB building and for performing all the fire drills in the morning when I am asleep.

I must thank everyone who knew me before I started the adventure of graduate school, these include my family and long-time friends: Kate McNenly, Peter McNenly, Casey McNenly, Michael Smith, Dawn Cooper, Mary McNenly, John Werle, Joseph Perisa, John Connolly, Stephanie Connolly, Kiet Tran, Ryan Knickerbocker, Patrick Lovelace, Michael Huber, Robert Doil and Jennie Doil. Through their support over the years, they have all had an impact on my life. I am a better person for knowing them and I am grateful that they still answer the phone when I call (except for Uncle Joe – who believes I am a telemarketer). I also need to express my deepest appreciation to my mother, who, with her sensitivity and finesse, always managed to ask me "when is your thesis going to be finished?" in a manner that would upset me the least.

Last, but most certainly not least, I must thank the love of my life, Anastacia. Her love, comfort, understanding and encouragement proved invaluable during this lengthy writing endeavor, and my gratitude to her is uncountably infinite. I would also like to apologize to Jürgen for having to bear the indignity of sharing such a small apartment with his two insufferable roommates.

TABLE OF CONTENTS

DEDICATION . . . . . **iii**

ACKNOWLEDGEMENTS . . . . . **iv**

LIST OF FIGURES . . . . . **xiii**

LIST OF TABLES . . . . . **xviii**

LIST OF APPENDICES . . . . . **xx**

CHAPTER

I. INTRODUCTION 1

1.1 Motivation 1
1.2 Problems facing the current simulation techniques for MEMS gas flows 5
1.2.1 Navier-Stokes simulation with slip model boundary conditions 7
1.2.2 Direct Simulation Monte Carlo 11
1.2.3 Boltzmann CFD 18
1.3 Objectives 19
1.4 Outline 24

II. EMPIRCAL CORRECTIONS TO THE NAVIER-STOKES SIMULATION 28

2.1 Background 31
2.1.1 Limitations to the slip model 34
2.2 Unified Models 38
2.2.1 Unified Poiseuille model 43
2.2.2 Unified Couette model 44
2.3 Investigative Method 45
2.3.1 Database construction 47
2.3.2 Optimum model coefficients 48

2.3.3 Unified model construction . . . 50
2.3.4 Predictive cases . . . 50
2.4 Least squares fit . . . 52
2.5 New Model Laws . . . 56
2.5.1 Non-linear data fitting . . . 57
2.5.2 Model sensitivity . . . 60
2.5.3 Candidate models . . . 64
2.5.4 Model Selection . . . 67
2.5.5 Model Error . . . 69
2.6 Empirical Model Prediction Performance . . . 73
2.6.1 Interpolation and Extrapolation . . . 74
2.6.2 Combination of Couette and Poiseuille Flow . . . 75
2.6.3 Tangential Momentum Accommodation Coefficient . 76
2.6.4 Helium Gas Flows . . . 78
2.6.5 Body force driven flow with uniform rates of suction and injection . . . 79
2.7 Summary . . . 80

III. QUASI-MONTE CARLO CONVERGENCE . . . 83

3.1 Integration Error . . . 85
3.1.1 Monte Carlo Integration Error . . . 86
3.1.2 Quasi-Monte Carlo Integration Error . . . 89
3.2 Discrepancy . . . 90
3.2.1 Calculation of the star discrepancy D_N^* . . . 94
3.2.2 Calculation of the extreme discrepancy D_N . . . 103
3.2.3 Calculation of the quadratic mean discrepancies T_N^* and T_N . . . 104
3.2.4 Calculation of the isotropic discrepancies H_N and J_N 106
3.3 Variation . . . 112
3.3.1 Variation in one dimension . . . 115
3.3.2 Variation in multiple dimensions . . . 120
3.4 Low discrepancy sequences versus optimal integration lattices 129

IV. LOW-DISCREPANCY SEQUENCES . . . 136

4.1 A Special Construction of the Weyl-Richtmyer Sequence . . . 139
4.2 Creating a Weyl-Richtmyer sequence with bounded continued fractions . . . 147
4.3 Sequence Implementation . . . 158
4.3.1 The pseudo-random sequence . . . 162
4.3.2 The Weyl-Richtmyer sequence . . . 163
4.3.3 The Halton sequence . . . 173
4.3.4 The Faure sequence . . . 177

4.3.5 The Niederreiter sequence in base 2 180

V. THE SIMULATION OF FREE MOLECULAR FLOW IN A TWO DIMENSIONAL DUCT 184

5.1 Basic Kinetics of Free Molecular Duct Flow 190
5.2 Markov Chain Simulation 201
5.3 Finite State Linear System Simulation 217
5.4 Nyström method 221
5.5 Particle Methods 236
5.5.1 Test Particle Monte Carlo Method 237
5.5.2 Absorption Weighted Monte Carlo Method 251
5.5.3 Quasi-Monte Carlo Method 263

VI. RESULTS FOR FREE MOLECULAR DUCT FLOW 270

6.1 The $L = 2$ Case 272
6.2 The $L = 5$ and $L = 10$ Cases 285
6.3 Duct Geometry Study ($0.5 \leq L \leq 10$) 292
6.4 Correlation between dimensions of the low-discrepancy sequences 311
6.4.1 Correlation between two dimensions of the Halton sequence 316
6.4.2 Correlation between two dimensions of the BCF-3 sequence 329
6.4.3 Correlation between two dimensions of the Faure sequence 335
6.4.4 Correlation between two dimensions of the Niederreiter sequence in base 2 339
6.4.5 The extent of the correlation present in the low-discrepancy sequences 348
6.5 Hybrid Quasi-Monte Carlo Simulation 352

VII. CONCLUSIONS 363

7.1 Summary 364
7.2 Future Work 375
7.2.1 Further improvements to the BCF-k sequences . . . 376
7.2.2 Free molecular flows with greater natural particle absorption 379
7.2.3 Reducing the dimension of the low-discrepancy sequences 381

APPENDICES **384**

BIBLIOGRAPHY **419**

LIST OF FIGURES

Figure

1.1 Examples of MEMS. 2

1.2 The range of physical validity of the current simulation methods for rarefied gas flows based on the Knudsen number. 6

1.3 Investigation path for the dissertation. 20

2.1 Poiseuille velocity profiles for $0.01 \leq Kn \leq 10$. 40

2.2 Couette velocity profiles for $0.01 \leq Kn \leq 10$. 41

2.3 Comparison of the continuum shear stress predicted by the unified model and by slip model alone. 43

2.4 Sensitivity of the model coefficients C_s and C_μ. 62

2.5 Comparison of the four non-linear models ability to capture the optimum slip coefficient for the Couette flow of argon gas. 66

2.6 Non-linear model construction for the optimum continuum corrections for argon gas flows. 68

2.7 L_2 error in the velocity profiles of the continuum corrections for Couette flow. 71

2.8 Relative error in the average shear stress τ_{xy} of the continuum corrections for Couette flow. 72

2.9 L_2 error in the velocity profiles of the continuum corrections for Poiseuille flow. 73

2.10 Combined Couette and Poiseuille flow for Argon gas at $Kn = 1$. . . 76

2.11 Poiseuille flow for Argon gas with different TMACs at $Kn = 1$. . . . 77

2.12 Poiseuille flow for different gases at $Kn = 1$ 78

2.13 Force driven duct flow with uniform suction and injection at the walls for Argon gas at $Kn = 1$. 80

3.1 Illustration of the possible supremum values in the calculation of the star discrepancy $D_N^*(P)$ of a two dimensional point set P. 95

3.2 Convergence of the approximate star-discrepancy $\overline{D}_N^*$ in two dimensions. 101

3.3 Examples of functions with bounded variation. 113

3.4 Examples of functions with unbounded variation. 118

3.5 Integration error of several test functions using a two dimensional Halton sequence. 124

3.6 Bounded variation in the sense of Vitali of two simple indicator function on $[0, 1)^2$. 125

3.7 Variation of an indicator function f on $[0, 1)^2$ using different domain partitions. 126

3.8 Illustration of the star discrepancy $D_N^*(P)$ convergence for different one-dimenisional point sets P. 131

4.1 The first 256 points of a two dimensional sequence. 138

4.2 Comparison of the constant in the bounding inequality for the extreme discrepancy $D_N \leq C_i N^{-1} \log(N + 1)$ for each dimension $1 \leq i \leq 255$ of different Weyl-Richtmyer sequences. 145

4.3 The number $\nu_k(T)$ of linearly independent quadratic surds which have a purely periodic continued fraction with coefficients bounded by k and a period less than or equal to T. 157

4.4 Comparison of the computation time needed to generate the pseudo-random sequence and the low-discrepancy sequences. 159

4.5 QMC performance comparison using different Weyl-Richtmyer low-discrepancy sequences. 170

4.6 The fraction of the components η of Halton sequence in s dimensions that can be calculated using the simple additive recursion. 177

5.1 Illustration of the two basic probability distributions used in the simulation of free molecular duct flow. 199

5.2 Illustration of the different types of transition probabilities for the gas molecules in the Markov chain simulation. 205

5.3 Convergence of the relative error in the Markov chain simulation of the conductance probability Ψ (for $L = 2$). 216

5.4 Convergence of the relative error in the finite-state linear system simulation of the conductance probability Ψ (for $L = 2$, 5 and 10). . 220

5.5 The Frobenius, L_1 , L_2, and L_∞ norms of the transition probability function $K(x, y)$ used to solve the conductance probability in a free molecular duct. 227

5.6 The probability $\overline{\varphi}_n$ of a particle remaining within the duct after $n+1$ wall collisions (for $L = 2$, 5 and 10). 230

5.7 Error convergence of the Nyström method using an n-point Gauss-Legendre rule to solve the conductance probability Ψ. 234

5.8 Comparison of the conductance probability Ψ calculated from the Nyström method and the approximation of Clausing. 236

5.9 The relative error of different particle simulations of the conductance probability Ψ (for $L = 2$). 244

5.10 Illustration of the non-physical molecular movement within the duct associated with a one dimensional low-discrepancy sequence. 247

5.11 Generating the sample trajectory for the absorption weighted (AW) method. 255

5.12 Distribution of the trajectory scores for the AWMC particle simulation of free molecular duct flows (for $L = 2$, 5 and 10). 259

5.13 Comparison of the test particle Monte Carlo simulation and the AWMC simulation (for $0.5 \leq L \leq 10$): (a) variance σ^2 in the trajectory scores; and (b) the approximate reduction factor for the relative error of the AWMC particle simulation. 260

5.14 Comparison of the test particle Monte Carlo simulation and the AWMC simulation (for $0.5 \leq L \leq 10$): (a) average number of particle moves per trajectory; and (b) the ratio of the total work (particle moves) needed by the test particle Monte Carlo simulation over the AWMC simulation to achieve the same error level. 262

5.15 Two dimensional projections of the trajectory score functions: (a) for the test particle simulation $S_{mc}(T(x_1, x_2))$; and (b) for the absorption weighted simulation $S_{awmc}(T(x_1, x_2))$. 266

5.16 Effect of the number of particle moves s used per sample trajectory in the QMC simulation on the overall error of the method. 268

6.1 Convergence of the relative error for the QMC particle simulation of the conductance probability Ψ (for $L = 2$). 279

6.2 Convergence of the relative error with respect to the computation time τ (in seconds) for the conductance probability Ψ (for $L = 2$). . 283

6.3 Convergence of the relative error for the QMC particle simulation of the conductance probability Ψ (for $L = 5$). 287

6.4 Convergence of the relative error for the QMC particle simulation of the conductance probability Ψ (for $L = 10$). 289

6.5 Convergence of the relative error after collecting 1, 4, and 16 ensembles for the QMC particle simulation using the Niederreiter sequence in base 2 (for $L = 10$). 297

6.6 The number of interior particle moves (and low-discrepancy sequence dimension s) used for the AWMC and QMC simulations. 298

6.7 The expected relative error of the particle simulations found after generating $N = 2^{23}$ sample trajectories. 299

6.8 The expected error convergence rate of the particle simulations. . . 302

6.9 The expected single sample error of the particle simulations. 303

6.10 The critical error E_{crit} of the QMC particle simulations. 305

6.11 Performance gains of the QMC particle simulations when compared to the expected error of the traditional test particle Monte Carlo method after $N = 2^{23}$ samples. 308

6.12 Plot of the two dimensional Halton sequence in prime bases $p_1 = 29$ and $p_2 = 31$. 317

6.13 Comparison of the running and local two dimensional correlation ρ_{12} for the Halton sequence $\mathbf{x}_n = (\chi_{29}(n), \chi_{31}(n))$. 322

6.14 The minimum sequence length N_{min} required for the Halton sequence constructed with prime bases p_1 and p_2 to be considered as uncorrelated as a random sequence. 324

6.15 Specific examples of the two dimensional construction patterns that can produce significant correlation between the dimensions of the low-discrepancy sequences. 331

6.16 Two dimensional construction patterns that appear after the first $N = 3844$ elements of the Faure sequence in base 31. 338

6.17 The extent of the two dimensional correlation present among the first 100 dimensions of the low-discrepancy sequences. 349

6.18 Convergence of the relative error for the hybrid QMC/MC simulation using different low-discrepancy sequences. 360

F.1 Example of a (t, m, s)-net in base 2, where $t = 1$, $m = 3$, and $s = 2$. 408

LIST OF TABLES

Table

2.1 TMAC values reported by Arkillic *et. al.* for various gas species in micro-machined silicon channels. 36

2.2 Table of the optimum coefficients C_s^* and C_μ^* for non-equilibrium Couette and Poiseuille flows. 55

2.3 Candidate non-linear model laws for the Couette flow slip coefficient for argon gas. 66

4.1 The first 10 irrational numbers used in the set **z** used to generate the BCF-3 low-discrepancy sequence. 155

6.1 The period of the near-cyclic construction patterns illustrated in Figure 6.17. 333

6.2 Comparison of the computation times of the hybrid QMC/MC simulation for different low-discrepancy sequences(for $L = 2$ and 10). . 361

7.1 The number of dimension pairs of the BCF-3 sequence which have significant two dimensional correlation which persists over sequence lengths comparable to the maximum used by the QMC particle simulations. 377

A.1 The first 16 points constructed for the van der Corput sequences in bases $b = 2$, 3 and 5. 387

B.1 The first 16 points constructed for a three dimensional Weyl sequence using the fractional parts of the irrational numbers $\sqrt{2}$, $\sqrt{3}$ and $\sqrt{5}$. 393

C.1 The first 16 points constructed for a three dimensional Halton sequence with prime bases $p_1 = 2$, $p_2 = 3$, and $p_3 = 5$. 396

D.1 The first 16 points constructed for a three dimensional Sobol' sequence using primitive polynomials over $\mathbb{F}_2$. 402

E.1 The first 16 points constructed for a three dimensional Faure sequence with a prime base $q = 3$. 406

F.1 The first 16 points constructed for a three dimensional Niederreiter sequence in base 3. 415

F.2 The value $t = t(s)$, for $1 \leq s \leq 20$, of different (t, s) sequences in base 2. 417

LIST OF APPENDICES

Appendix

A. The van der Corput sequence (1935) 385

B. The Weyl-Richtmyer Sequence (1916/1951) 389

C. The Halton Sequence (1960) 394

D. The Sobol' Sequence (1967) 397

E. The Faure Sequence (1982) 403

F. The Niederreiter (t, s)-Sequence (1987) 407

CHAPTER I

INTRODUCTION

1.1 Motivation

Micro-electro-mechanical systems, or MEMS, represent a $5.3 billion industry (in 2003) that is expected to grow to nearly $10 billion by 2008.[1] Current MEMS applications are found across a wide range of scientific and engineering fields, which include: automotive, aerospace, biology, chemistry, computer science, medicine, optics, and telecommunications. To illustrate the length scales of current MEMS design, four devices developed at Sandia National Laboratories[2] are shown in Figure 1.1. Creating devices with these length scales is particularly challenging because designs that are successful at the human-scale, or meso-scale $\mathcal{O}\,(\mathrm{m})$, often do not function as intended at the micro-scale $\mathcal{O}\,(\mu\mathrm{m})$. One reason is that the surface area to volume ratio increases as the characteristic length scales shrink; and as a consequence, phenomena which are dependent on the surface area of the system (*e.g.* friction) become more important at the micro-scale. As the length scales shrink further in MEMS design and begin to approach the nano-scale $\mathcal{O}\,(\mathrm{nm})$, quantum mechanics must also be considered in some applications. The rapid growth of the industry has created many

[1] See the July 16, 2004 online edition of *Small Times* magazine (`www.smalltimes.com`).

[2] Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

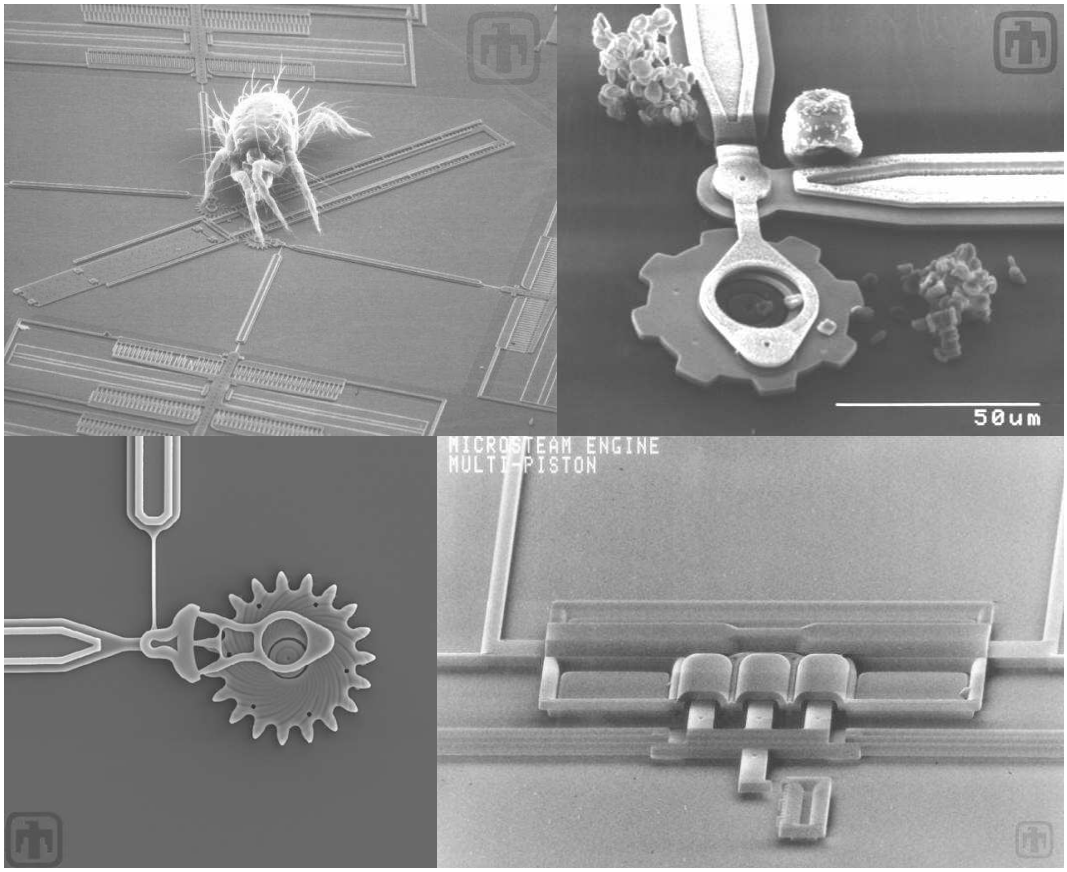

Figure 1.1: Examples of MEMS (clockwise from top-left): a spider mite dwarfing three linear comb drives used for the actuation of a micro-mirror system; a collection of red blood cells and a single grain of pollen next to a micro-gear; a three piston micro-scale steam engine; and an air-cushioned bearing for a micro-gear. All images courtesy of Sandia National Laboratories, SUMMiT^{TM} Technologies, `www.mems.sandia.gov`.

active areas of research, which are focused on improving the theoretical analysis and computer simulations required for all aspects of MEMS design: aerodynamic, electro-magnetic, mechanical, optical, structural, and thermal.

The term fluidic MEMS refers to any device in which the control or distribution of a fluid flow are important (see [51, 67, 69] for an overview). Moreover, fluidic MEMS that are designed to control gas flows are of particular interest to this investigation. The intended purpose of a MEMS device does need to actively involve gas flows for the aerodynamic effects to be critical to the design. For example, any MEMS device with moving parts that is not vacuum-sealed is likely to be affected by air-friction. In fact, a mechanical system which functions perfectly at the meso-scale may be rendered inoperable when shrunk to the micro-scale because the relative magnitude of air-friction grows with the surface area to volume ratio. The accuracy of the boundary conditions, therefore, become very important in any analysis of simulation of micro-scale gas flows. There are additional challenges facing aerodynamic analysis at the micro-scale because common meso-scale assumptions and solution techniques are no longer valid. In the case of low-speed micro-channel gas flows with a Mach number $M < 0.1$, the assumption of incompressibility is not physically accurate as there is a significant drop in pressure along the length of the channel, which leads to compressibility effects (see the experimental results of [4, 6, 5, 170, 145]). Thus, the aim of the research presented in this investigation is to improve the available simulation methods for general micro-scale flows, in order to better evaluate fluidic MEMS designs.

In addition to aforementioned micro-channels, some designs of micro-scale pumps and valves have also been tested experimentally (see [27, 29, 122, 177]). These basic construction elements for controlling MEMS flows have been used successfully to de-

velop lab-on-chip MEMS, which manipulate flow for testing processes that demand a high degree of sensitivity. Lab-on-chips are also able to combine a microprocessor on the same silicon structure as the MEMS device, in order to process and analyze the data as it is collected. This has the potential to greatly increase the throughput of the testing process compared to the traditional meso-scale methods. Some specific applications of lab-on-chips, which have recently been developed and tested experimentally, include: (i) fluid density and chemical concentration measurements [169]; (ii) DNA analysis [53, 132]; and (iii) detection of explosive particles at the nanogram level [137]. Successful fluidic MEMS application are not solely limited to internal flows and lab-on-chips, micro-scale thrusters have also recently been demonstrated. In particular, MEMS thrusters have been developed and tested experimentally for applications involving micro-scale locomotion under standard atmospheric conditions [141], and precise satellite attitude control in the space environment [64].

While the list of successful fluidic MEMS devices is significant, the future applications currently being developed are even more impressive. One of the main goals of fluidic MEMS research is to develop an atmospheric micro-scale flier. Such a device clearly has numerous security applications as it offers nearly undetectable reconnaissance. Further, its low manufacturing cost would allow large arrays to be deployed for the detection and tracking of minute amounts of air-borne toxins or radiation. Current development of an atmospheric micro-scale flier is especially difficult considering that almost all of the fluidic MEMS designs tested to-date involve internal flows. Fortunately, the recent efforts of Martin and Boyd [108] aim to fill this void in the MEMS research by developing both a micro-scale wind tunnel, and micro-scale airfoils to test within the facility. The development of micro-scale fliers is also important in the bio-medical field, where such devices have the potential to detect,

diagnose, and treat at levels unparalleled by any of the current methods found in medicine.

The cost of developing and testing any MEMS prototype is appreciable; and thus, the numerical simulation of MEMS becomes an extremely important cost-savings tool in the design of a new device. Unfortunately, for many fluidic MEMS applications involving gas flows, there does not exist a simulation technique which is both accurate and computationally efficient. Continuum-based methods, such as the Navier-Stokes simulation, have a relatively low computational cost, but are not physically accurate for a wide range of flow conditions commonly found in fluidic MEMS. On the other hand, particle-based methods, such as direct simulation Monte Carlo, are physically accurate for fluidic MEMS, but suffer from exceedingly large computational costs that are difficult to manage unless one has a supercomputer available. Given the apparent lack of an accurate and efficient simulation technique for micro-scale flows, the goal of this investigation is develop new approaches in an effort to improve the design analysis for fluidic MEMS applications.

1.2 Problems facing the current simulation techniques for MEMS gas flows

The two most widely used simulation techniques for MEMS gas flows are the Navier-Stokes simulation with slip model boundary conditions, and the Direct Simulation Monte Carlo (DSMC) method of Bird [16]. The suitability of a particular simulation technique depends on both the average flow speed and the amount of rarefaction, or non-equilibrium, present in the MEMS gas flow. The amount of rarefaction is most often expressed by the non-dimensional Knudsen number Kn, which is defined as the ratio of the mean free path ℓ_p to the characteristic length scale L

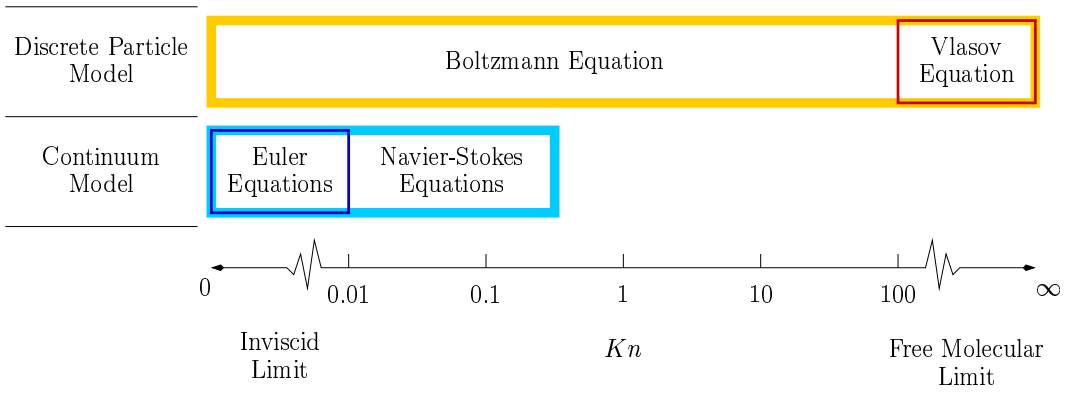

Figure 1.2: The range of physical validity of the current simulation methods for rarefied gas flows based on the Knudsen number.

(*i.e.* $Kn = \ell_p/L$).[3] As the Knudsen number increases, the number of inter-molecular collisions occurring within the volume L^3 decreases. Since the collisions between the gas molecules drive the flow toward thermodynamic equilibrium (or equivalently, statistical equilibrium), a flow is said to have deviated further from equilibrium as the Kn increases. Continuum methods, such as the Euler and Navier-Stokes equations, are based on the assumption that the gas flow is in local thermodynamic equilibrium; as a consequence, these methods are only physically accurate when $Kn \ll 1$. For larger Knudsen numbers (*i.e.* $Kn = \mathcal{O}(1)$ and greater), the actual statistical behavior of the gas molecules must be simulated in order to achieve a physically valid solution. Discrete particle models based on the Boltzmann equation (*e.g.* DSMC) or the Liouville equation (*e.g.* Molecular Dynamics MD) are therefore used for these non-equilibrium gas flows. It is important to note that these discrete particle models are valid for all Knudsen numbers.

For comparison, the range of Knudsen numbers over which the simulation methods are considered accurate is given in Figure 1.2 (see [16]). Note that the Vlasov equation is equivalent to Boltzmann equation for gas flows when there are no inter-

[3]The mean free path ℓ_p is defined as the distanced traveled, on average, by a gas molecule before it collides with another gas molecule.

molecular collisions present. A large number of fluidic MEMS applications have low-speed gas flows operate within a range of Knudsen numbers termed the transition regime. Although there is not an exact definition, the lower bound of the transition regime is typically assumed to be in the range of 0.01 to 0.1 while the upper bound is assumed to be in the range 10 to 100. The problem with the continuum-based Navier-Stokes simulation of fluidic MEMS is that it is not physically accurate for much of the transition regime, as illustrated in Figure 1.2. In contrast, the particle-based DSMC method is physically accurate throughout the transition regime, but it has a very high computation cost when used to simulate low-speed flows. These two simulation techniques are briefly discussed in the remainder of this section with regard to the difficulties they encounter for MEMS gas flows.

1.2.1 Navier-Stokes simulation with slip model boundary conditions

Strictly speaking, the Navier-Stokes equation(s) refers only to the conservation of momentum (in three dimensions) under the assumption of a Newtonian shear stress closure [187]. The term Navier-Stokes simulation, however, is used in practice to describe the numerical approximation of the entire system of conservation laws, which govern the continuum dynamics of a viscous flow. These include the following equations for the conservation of mass (1.1), momentum (1.2), and energy (1.3). Specifically,

$$\frac{\partial \rho}{\partial t} + \nabla(\rho \mathbf{v}) = 0 \tag{1.1}$$

$$\frac{\partial}{\partial t}\rho\mathbf{v} + \vec{\nabla} \cdot \left(\rho \mathbf{v} \otimes \mathbf{v} + pI - \overline{\overline{\tau}}\right) = \rho \mathbf{g} \tag{1.2}$$

$$\frac{\partial}{\partial t}\rho(e + \tfrac{1}{2}|\mathbf{v}|^2) + \nabla\left(\rho\mathbf{v}(h + \tfrac{1}{2}|\mathbf{v}|^2) - \alpha\nabla T - \overline{\overline{\tau}} \cdot \mathbf{v}\right) = \rho\mathbf{g} \cdot \mathbf{v} + Q_H, \tag{1.3}$$

where $\otimes$ denotes the outer tensor product of a vector, I denotes the identity tensor, and the following quantities are specified: the external body force $\mathbf{g}$, the thermal

conductivity coefficient α, and the external heat source Q_H. The system (1.1-1.3) represents five partial differential equations governing the evolution of five macroscopic flow quantities in time t and space $\mathbf{x}$: (i) the density $\rho(\mathbf{x}, t)$; (ii) the velocity vector $\mathbf{v}(\mathbf{x}, t) = (v_1, v_2, v_3)$; and (iii) the pressure $p(\mathbf{x}, t)$. To close this system of equations, the remaining thermodynamic quantities (e - internal energy, h - enthalpy, and T - temperature) are represented in terms of the density and pressure using the ideal gas law. Further, the shear stress tensor $\overline{\overline{\tau}}$ is calculated in terms of the velocity gradients assuming the Newtonian shear stress closure with the Stokes relation, which yields (using Einstein summation)

$$\overline{\overline{\tau}} = \tau_{ij} = \mu \left(\frac{\partial v_j}{\partial x_i} + \frac{\partial v_i}{\partial x_j} - \frac{2}{3} \frac{\partial v_k}{\partial x_k} \delta_{ij} \right),$$

where μ is the coefficient of viscosity, which is specified, and δ_{ij} is the Kronecker delta. A thorough review of the Navier-Stokes simulation is given by Hirsch in [63].

At the solid boundary surfaces in the flow domain, the velocity field $\mathbf{v}(\mathbf{x}, t)$ in the Navier-Stokes simulation is most often treated by no-slip boundary conditions, which are physically consistent in the continuum limit (*i.e.* $Kn \to 0$). Specifically, $\mathbf{v}(\mathbf{x}, t) = (0, 0, 0)$ for all points $\mathbf{x}$ on the solid boundary surface.[4] As a gas flow begins to deviate slightly from local thermodynamic equilibrium (*i.e.* $0 < Kn \ll 1$), the first non-equilibrium regions develop near the solid boundary surfaces. These non-equilibrium regions are referred to as the Knudsen layer, and typically have an approximate thickness of several mean free paths. The gas molecules in the Knudsen layer do not undergo a sufficient number of collisions to reach local thermodynamic equilibrium with the solid boundaries. As a consequence, there can exist a non-zero tangential velocity at the boundary, which is termed the slip velocity. When the

[4]This is assuming the solid boundary surface is stationary. More generally, $\mathbf{v}(\mathbf{x}, t) = \mathbf{w}(\mathbf{x}, t)$ for all points $\mathbf{x}$ on the solid boundary where $\mathbf{w}(\mathbf{x}, t)$ is the local velocity of the solid surface boundary.

Knudsen layer is sufficiently small, almost all regions of the flow are still accurately represented by the Navier-Stokes simulation except at the solid boundaries. It is therefore entirely reasonable to expect that the Navier-Stokes simulation still yields an accurate approximation when Kn is sufficiently small provided that the no-slip boundary conditions are corrected. The purpose of a slip model is then to correct the boundary conditions of the Navier-Stokes simulation by estimating the non-zero slip velocity using known macroscopic flow quantities from the equilibrium regions.

The concept of a slip model dates back to the origins of gas kinetic theory, with Maxwell proposing the first slip model in [109]. Maxwell derived the slip model using a perturbation analysis of the behavior of the velocity distribution function within a mean free path of a solid boundary. From this analysis (assuming a fully diffuse wall), the tangential slip velocity u_s at the wall is given by

$$u_s = \frac{2}{3}\ell_p \left.\frac{\partial u}{\partial n}\right|_{wall},$$

where ℓ_p is the mean free path. Note that Maxwell's slip model recovers the no-slip boundary condition in the continuum limit $Kn = 0$. More importantly though, Maxwell's slip model is mathematically consistent for non-equilibrium gas flows (with small Kn) during the approach of the limit $Kn \to 0$ because it is derived directly from kinetic theory valid at all Knudsen numbers.

Although non-equilibrium temperature effects are not considered in this investigation, it should be noted that the lack of collisions within the Knudsen layer similarly prevents the gas temperature from reaching equilibrium with the solid boundary. Thus, there may also exist a non-zero temperature jump at the solid boundaries. Based on the same type of perturbation analysis as Maxwell's slip model, von Smoluchowski [180] derives a similar model to estimate the temperature jump at the solid

boundaries. During the century that has passed since the models of Maxwell and von Smoluchowski were first published, many new models have been proposed to yield more accurate corrections to the continuum-based boundary conditions under certain flow conditions (see for example [11, 12, 15, 25, 39, 69, 80, 111, 112, 138]). All of these new models, however, still retain the same basic structure as the original models of Maxwell and von Smoluchowski.

The Navier-Stokes simulations with slip model boundary conditions have been successfully implemented by Cai *et. al.* [24] for micro-channel flows, and by Sun *et. al.* [171] for micro-scale airfoils. The results in [24, 171] demonstrate the limited range of physical accuracy of the Navier-Stokes simulation in the transition regime, which is the main drawback of the method for fluidic MEMS simulations. While the exact range of applicability depends on the flow geometry and desired accuracy, the Navier-Stokes simulation with slip model boundary conditions is widely considered to be physically accurate when $Kn \lessapprox 0.1$. At $Kn \approx 0.1$, the Knudsen layer in a micro-channel easily covers most of the flow domain. As the Knudsen number increases beyond $Kn \approx 0.1$, it is unreasonable to expect that modifying the continuum-based boundary condition is able to properly account for the non-equilibrium effects, which occur everywhere within the flow. Because the Navier-Stokes simulation with slip model boundary conditions is only physically accurate for near-continuum non-equilibrium gas flows, the method is not particularly well-suited to simulate future fluidic MEMS applications. Following the development of micro-electronics for the computing industry, there is a similar drive to produce smaller devices throughout the MEMS industry in an effort to improve their response time and sensitivity, while lowering their manufacturing costs. As the size of MEMS decreases, so too does the characteristic length scales in the gas flows found in the fluidic MEMS applications,

which, in turn, increases the Knudsen number. Thus, a Navier-Stokes simulation that produces a good solution for a current fluidic MEMS design may not be sufficiently accurate to evaluate the next generation of the design.

1.2.2 Direct Simulation Monte Carlo

The Boltzmann equation is a single non-linear integro-differential equation in up to 7 dimensions, and provides an accurate description [5] of gas flows for all Knudsen numbers, as shown in Figure 1.2. Let $F = F(\mathbf{x}, \mathbf{u}, t)$ denote the number of gas molecules located in the infinitesimal volume of space at $\mathbf{x} = (x_1, x_2, x_3)$ which travel with a velocity in the infinitesimal neighborhood of $\mathbf{u} = (u_1, u_2, u_3)$ at time t. The function F is termed the velocity distribution function and its evolution is governed by the Boltzmann equation (see [16, 25, 54, 80, 179]). Specifically,

$$\frac{\partial F}{\partial t} + \mathbf{u} \cdot \nabla_{\mathbf{x}} F + \mathbf{g} \cdot \nabla_{\mathbf{u}} F = \iiint_{-\infty}^{\infty} \int_0^{2\pi} \int_0^{\pi} \varpi \sin\chi \big(F' F_2' - F F_2\big) S(\varpi, \chi) d\chi d\varepsilon d\mathbf{v}, \quad (1.4)$$

where $F_2 = F(\mathbf{x}, \mathbf{v}, t)$ is the velocity distribution function of a collision partner traveling at a velocity $\mathbf{v}$, $F' = F(\mathbf{x}, \mathbf{u}', t)$ and $F_2' = F(\mathbf{x}, \mathbf{v}', t)$ with the primes denoting the post-collision velocities (or distributions of velocities) of two molecules, $\mathbf{g}$ is an accelerative body force, (ε, χ) are the two trajectory angles characterizing a binary collision, $\varpi = |\mathbf{u} - \mathbf{v}|$ is the relative speed between two colliding molecules, and $S(\varpi, \chi)$ is the differential cross-section of the collision based on the inter-molecular forces.

[5] The Boltzmann equation is obtained from the more general Liouville equation under the following assumptions (see [179] p. 333): (i) the range of inter-molecular force is much smaller than the average distance between collisions (*i.e.* only binary collisions occur); (ii) there is no correlation between the initial velocities of two molecules undergoing a collision (molecular chaos/irreversibility); and (iii) the distribution function does not vary appreciably over a distance (or time) on the order of the range of inter-molecular forces (or duration of a collision).

The Boltzmann equation (1.4) must have prescribed initial and boundary conditions in order to be well-posed. Let $\mathcal{V} \subset \mathbb{R}^3$ denote an arbitrary three dimensional flow domain with a boundary surface $\mathcal{S}$. The initial condition for the velocity distribution function F is then well-posed if, at some time t_0,

$$F(\mathbf{x}, \mathbf{u}, t_0) = F_0(\mathbf{x}, \mathbf{u}),$$

where $F_0 = F_0(\mathbf{x}, \mathbf{u})$ is a known non-negative function defined at every $\mathbf{x} \in \mathcal{V}$ and every $\mathbf{u} \in \mathbb{R}^3$. In most applications of the Boltzmann equation to non-equilibrium gas flows, there are two basic boundary conditions: (i) the inflow/outflow-type; and (ii) the surface interaction/reflection-type. Let $\mathcal{I} \subset \mathcal{S}$ denote the regions of the boundary surface where there is an inflow/outflow type boundary condition, and further define $\hat{\mathbf{n}}(\mathbf{x})$ as the unit surface normal (pointing into $\mathcal{V}$) for all $\mathbf{x} \in \mathcal{S}$. The inflow/outflow-type boundary condition for the velocity distribution function F is then well-posed if

$$F(\mathbf{x}, \mathbf{u}, t) = F_i(\mathbf{x}, \mathbf{u}, t)$$

where $F_i = F_i(\mathbf{x}, \mathbf{u}, t)$ is a known non-negative function defined (for all times $t > t_0$) at every $\mathbf{x} \in \mathcal{I}$ and every $\mathbf{u} \in \mathbb{R}^3$ such that $\mathbf{u} \cdot \hat{\mathbf{n}}(\mathbf{x}) > 0$.

Similarly, let $\mathcal{W} \subset \mathcal{S}$ denote the regions of the boundary surface where there is an interaction/reflection-type boundary condition. This type of boundary condition is then well-posed if

$$F(\mathbf{x}, \mathbf{u}, t) = \int_{\mathbf{v} \cdot \hat{\mathbf{n}}(\mathbf{x}) < 0} F(\mathbf{x}, \mathbf{v}, t) K(\mathbf{u}, \mathbf{v}) d\mathbf{v},$$

where $K(\mathbf{u}, \mathbf{v})$ is a non-negative function defined for all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^3$ that satisfy $\mathbf{u} \cdot \hat{\mathbf{n}}(\mathbf{x}) > 0$ and $\mathbf{v} \cdot \hat{\mathbf{n}}(\mathbf{x}) < 0$. The function $K(\mathbf{u}, \mathbf{v})$ is often referred to as the scattering kernel. In particular, it represents the probability that a molecule intersects the

boundary $\mathcal{S}$ with a velocity $\mathbf{v}$, from within the domain $\mathcal{V}$, and is then reflected back into the domain with a velocity $\mathbf{u}$. The scattering kernel $K(\mathbf{u}, \mathbf{v})$ may also be generalized to include variations in both time and space. When the interaction/reflection-type boundary condition is used to represent a solid wall with no net absorption of gas molecules at the surface, the scattering kernel must also satisfy an additional normalization condition to be physically consistent. The boundary conditions are discussed in greater detail for the Boltzmann equation in the monographs of Cercignani [25] and Kogan [80], while specific details of their implementation in the DSMC method are given in the monograph of Bird [16].

The DSMC method approximates the Boltzmann equation (1.4) by indirectly simulating the stochastic molecular behavior governed by the equation, rather than using standard discretization techniques to directly solve the complicated non-linear integro-differential equation. In all DSMC simulations (except at the nano-scale), the number of molecules is simply too large to track their trajectories individually. As a consequence, a relatively small number of "simulated particles" are used in practice, where each simulated particle represents the local behavior of a large number of real gas molecules. The location and velocity of these simulated particles are then tracked at discrete times as they travel throughout the flow geometry. To calculate the spatial variation of the velocity distribution function F in (1.4), the DSMC method divides the flow geometry into discrete computational cells. The cell size is selected such that the variation in the velocity distribution function F is relatively small, which requires all dimensions of the cell to be smaller than the local mean free path ℓ_p.[6] Since the simulated particles evolve according to the Boltzmann equation (at least in a probabilistic sense), they provide a snapshot of the local velocity distribution

[6]The cell dimensions are generally selected to be smaller than $\ell_p/3$ in order to maintain physical accuracy in the DSMC method (see [16]).

function F when sorted into the computational cells. The expected behavior of the non-equilibrium gas flows is then approximated in the DSMC method by collecting a large number of these representative snapshots in each computational cell.

In the discrete time evolution of the system, the location and velocity of the simulated particles are updated after each time step. To simplify this update calculation, the time step is chosen to be a fraction (typically less than $\frac{1}{3}$) of the mean time between collisions; and as a consequence, only a fraction of the simulated particles undergo a collision during this time. More importantly, the fraction of real molecules colliding multiple times during the time step is sufficiently small so as to be considered negligible in practice. Therefore, the advection and collision processes are able to be accurately updated independently for the simulated particles by limiting the time step in this manner, which is equivalent to the physics-splitting found in traditional CFD methods. The advection update[7] of each simulated particle calculates the new particle location and velocity based on its ballistic trajectory. In contrast, the enforcement of the boundary conditions and the collision update[8] are probabilistic in nature for the simulated particles. To obtain an accurate statistical representation of the random collision process, the number of real molecules represented by each simulated particle is generally selected such that each computational cell contains at least 20 to 30 simulated particles on average (see Bird [16]). Because some steps of the update calculation are random, the DSMC method (as its name indicates) is thus considered a Monte Carlo method.

The DSMC method is, in essence, a collection of simulated particles that behave as if they were sampled directly from the local velocity distribution function as

[7]That is, the time integration of the gradient terms on the left-hand side of the Boltzmann equation in (1.4).

[8]That is, the time integration of the integral collision operator on the right-hand side of the Boltzmann equation in (1.4).

governed by the Boltzmann equation (1.4). The distribution of simulated particles at any instant is not, however, an accurate representation of the expected velocity distribution function because there is only a small number of particles in each cell. When all the simulated particles in a cell are viewed as a single set, some of the new velocities and locations in the set will be randomly generated during the update of each time step. Under the assumption of steady-flow, the distribution of simulated particles after each time step can thus be considered an independent realization, or a sample, of the true velocity distribution.[9] By averaging together over many time steps the distribution of simulated particles in each computational cell, the DSMC method is able to accurately represent the expected velocity distribution function. In fact, the Central Limit Theorem [47] implies that the average distribution of simulated particles in the DSMC method is expected to converge (in a probabilistic sense) at a rate of $\mathcal{O}\left(N^{-1/2}\right)$ to the true velocity distribution function governed by the Boltzmann equation (1.4).

The overwhelming majority of DSMC simulations of non-equilibrium gas flows do not actually need all the detailed microscopic information contained within the velocity distribution function. Usually, the goal of most simulations is simply to obtain accurate approximations to the macroscopic properties of the flow, which include: the density, average flow velocity, pressure, shear stress, and temperature. These macroscopic properties are determined by taking the appropriate moments of the velocity distribution function. For example, the density of the flow $\rho(\mathbf{x})$ is determined by the zeroth-order moment of the velocity distribution function $F =$

[9] It is suggested by some (*e.g.* Bird [16]) that the distribution of simulated particles in each cell should only be considered independent after a sufficient number of time steps have elapsed in order to allow for all the particles, on average, to undergo at least one collision. The frequency at which the simulated particles are sampled in each cell does not, however, affect the actual convergence to the true velocity distribution function. Hence, it is still acceptable to sample the DSMC simulation after each time step.

$F(\mathbf{x}, \mathbf{u})$, which yields

$$\rho(\mathbf{x}) = \int_{-\infty}^{\infty} mF(\mathbf{x}, \mathbf{u}) d\mathbf{u},$$

where m is the molecular mass of the species. The average velocity of the flow $\mathbf{v}(\mathbf{x}) = (v_1, v_2, v_3)$ is determined by the first-order moment of the velocity distribution function, and is given by

$$v_i(\mathbf{x}) = \frac{1}{\rho(\mathbf{x})} \int_{-\infty}^{\infty} mu_i F(\mathbf{x}, \mathbf{u}) d\mathbf{u} \quad \text{for } i = 1, 2, 3.$$

Likewise, the pressure, shear stress, and temperature of the flow are determined by the second-order moments of the velocity distribution function. With respect to the memory requirements, it is much more computationally efficient to only store the running averages, or tallies, for the velocity moments of interest rather than storing the entire approximation to the velocity distribution function. A thorough review of all the implementation details of the DSMC method, including a complete FORTRAN code with all the necessary algorithms, is provided by Bird in [16].

There are two common implementations of DSMC for the approximation of the Boltzmann equation: (i) the method of Bird [16]; and (ii) the method of Nanbu [121]. The only significant difference between the two methods occurs within the approximation of the collision integral operator of the Boltzmann equation (1.4). If N_p is the number of simulated particles in a computational cell, then the operation cost of computing the particle collisions is $\mathcal{O}\left(N_p \log N_p\right)$ for the method of Bird and $\mathcal{O}\left(N_p^2\right)$ for the method of Nanbu. Further, the method of Bird conserves energy during each binary collision, while the method of Nanbu only conserves energy in an average sense and thus the total system energy follows a random walk. Due to its computational efficiency and the fact that energy is conserved in-detail, the DSMC method of Bird [16] is more often used in practice. Note that the term "DSMC," as

it relates to a specific simulation technique, refers to the DSMC method of Bird in this investigation, unless specifically stated otherwise. The method of Nanbu does, however, offer some advantages with respect to the more rigorous mathematical study of the convergence of particle simulation to the Boltzmann equation, as noted in [8, 10, 86, 87, 88]. This is due in part to the Nanbu method being derived directly from the Boltzmann equation, which makes the task of establishing the consistency of the Nanbu method easier. It should be noted, however, that both DSMC methods are now known to yield mathematically consistent approximations to the Boltzmann equation. In particular, Babovsky and Illner originally proved the consistency of the method of Nanbu in [10], while Wagner more recently proved the consistency of the method of Bird in [182].

The main drawback to the DSMC method is the relatively slow convergence of the moments of the velocity distribution function collected from the simulated particles in each cell. The convergence rate of DSMC, as with all Monte Carlo techniques, is $\mathcal{O}\left(N^{-1/2}\right)$, where N is the number of independent samples; however, the convergence rate alone does not determine the total computational cost of DSMC. The magnitude of the natural statistical fluctuations present in the method relative to the average bulk velocity of the flow (or equivalently, the desired error level) is important as well. Specifically, the computational cost of DSMC increases quadratically as the average bulk velocity, or desired error level, decreases. The DSMC method is thus best-suited for the simulation of high-speed non-equilibrium gas flows such as the hypersonic flows associated with atmospheric re-entry vehicles.

To better illustrate the computational cost of the DSMC method, consider the simulation of the following two free stream flows of nitrogen gas at STP[10]: (i) at 1,000

[10]Standard temperature and pressure (STP) is 0°C at 100.0 kPa as defined by the International Union of Pure and Applied Chemistry (IUPAC).

m/sec (approximately Mach 3); and (ii) at 1 m/sec. The mean molecular speed of nitrogen gas at STP is $\overline{v}_{N_2} = 455$ m/sec with a standard deviation $\sigma_{N_2} = 285$ m/sec. Suppose one wants to resolve the average velocity of these two free stream flows to an accuracy of 1% (with a 95% confidence interval) using the DSMC method. Then, by the Central Limit Theorem [47], resolving the 1,000 m/sec flow requires around 1,500 independent samples of the velocity distribution function to reach the desired error level. In contrast, the 1 m/sec flow requires more than 1.5 *billion* independent samples to be simulated in the DSMC method in order to achieve the same 1% accuracy. The magnitude of the statistical fluctuations in molecular velocities represents the thermal energy of the flow and is independent of the characteristic length scales of the flow geometry. Consequently, the DSMC method is expected to suffer from extremely long computation times when simulating the low-speed non-equilibrium gas flows commonly found in fluidic MEMS applications.

1.2.3 Boltzmann CFD

There is also a third simulation technique that has been developed for non-equilibrium gas flows referred to as Boltzmann CFD (Computational Fluid Dynamics); however, it is much less popular than the slip-corrected Navier-Stokes simulation and the DSMC method. Boltzmann CFD uses the discretization techniques commonly found in CFD[11] to obtain a consistent numerical approximation to the Boltzmann equation (1.4). Working together, Ohwada, Sone, and Aoki [134, 133, 164, 163, 165] have developed Boltzmann CFD simulations for several different one dimensional gas flows assuming a linearized collision operator for hard-sphere molecules. Each spatial and temporal point in the Boltzmann CFD simulation

[11] For example, partial derivatives may be approximated using a finite difference, finite element, or finite volume approach, while integral terms may be approximated using a Newton-Cotes integral rule or Gaussian quadrature.

requires a discretization of the velocity phase space in order to represent the velocity distribution function governed by (1.4). As a consequence, the simulation of a one dimensional flow geometry with Boltzmann CFD requires a minimum of three dimensions, even under the assumption of a steady-state solution. More general two and three dimensional flow geometries require the full three dimensional velocity space to be calculated and stored. Boltzmann CFD must therefore simulate a total of five or six dimensions in these cases. Further, each time step in Boltzmann CFD requires an evaluation of the integral collision operator, which for the update of each simulated point in the velocity phase space requires, at the very minimum, a direct summation over all the simulated points in the local phase space. The main drawback to Boltzmann CFD is thus the high computational cost in terms of both memory and simulation time of the method, which is the reason why its use is infrequent. These dimension problems become even more difficult to overcome in simulations involving: (i) a large range of particle velocities (*e.g.* hypersonic flows); (ii) collisions with additional degrees of freedom (*e.g.* poly-atomic molecules and high temperature flows); and (iii) multiple gas species.

1.3 Objectives

The accurate and efficient simulation of low-speed non-equilibrium gas flows is a much sought, yet elusive, goal in fluidic Micro-Electro-Mechanical-Systems (MEMS) research. The two most popular simulation techniques for MEMS applications involving gas flows are the Navier-Stokes simulation with slip boundary conditions and the direct simulation Monte Carlo (DSMC) method of Bird. In almost every application, the Navier-Stokes simulation converges to a solution much faster than the DSMC method. The Navier-Stokes simulation, however, is only physically accurate

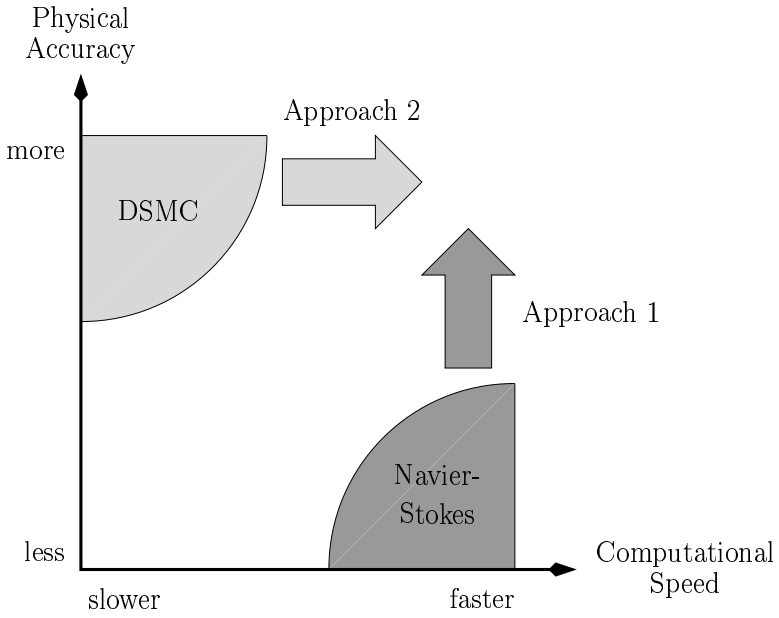

Figure 1.3: Investigation path for the dissertation.

for near-equilibrium gas flows; that is, when the Knudsen number $Kn \lessapprox 0.1$. The DSMC method, in contrast, is physically valid for the entire range of Knudsen numbers $0 < Kn < \infty$, but suffers from a relatively slow convergence rate $\mathcal{O}\left(N^{-1/2}\right)$ (where N is the number of samples). As previously noted, when the average bulk velocity is significantly slower than the average speed of the simulated particles, which is common in many fluidic MEMS, the problems associated with the slow convergence result in substantial computation cost. In fact, when the average velocity in the fluidic MEMS is on the order of [mm/sec], the computation time of the DSMC method is often intractably long on all but the world's largest supercomputers.

To confront these challenges facing the popular simulation techniques, two approaches are developed in this investigation in an effort to achieve an accurate and efficient simulation of low-speed non-equilibrium gas flows. The general goal of each approach is illustrated in Figure 1.3, which plots the relative speed and accuracy of the major methods under consideration. The first approach aims to extend the ac-

curacy of the computationally efficient Navier-Stokes simulation to higher Knudsen numbers in the transition regime by introducing empirical corrections to both the boundary conditions and transport closures of the method. Similar empirical corrections have been introduced by Karniadakis and Beskok [15] and Bahukudumbi *et. al.* [12], and thus the focus in this investigation is to better establish the range of applicability of such methods. The second, and more ambitious, approach is to develop a quasi-Monte Carlo (QMC) particle simulation that retains the physical accuracy of the DSMC method while converging at a rate faster than $\mathcal{O}\left(N^{-1/2}\right)$. The QMC method refers to any integral approximation that achieves a near-linear theoretical error convergence rate $\mathcal{O}\left(N^{-1+\epsilon}\right)$, for all $\epsilon > 0$, when sampled by a set of points in a manner similar to the Monte Carlo method. The key difference is that the set of sample points for the QMC method are not generated at random, instead they are deterministically selected in order to obtain the most uniform distribution possible. The design of a successful QMC method is not, however, a trivial undertaking. Even for relatively simple problems, great care must be exercised when developing each step of the simulation process or else the near-linear convergence rate is not attained in practice. In fact, it is the understanding of the author that no QMC particle simulation for general non-equilibrium flows has ever demonstrated near-linear convergence. Given the difficulty associated with the task, the QMC particle simulation is only developed in this investigation for free molecular, or collision-less flows.

The two approaches developed in this investigation are not the only techniques under consideration at this time, for improving the simulation of non-equilibrium gas flows. There are, in fact, several alternative approaches currently being researched that are noteworthy and deserve mention here. One of these alternative approaches involves the reduction of the statistical scatter that is naturally present among the

simulated particles. If the statistical scatter is smaller, then variance of the Monte Carlo method is also reduced, which, by the Central Limit Theorem [47], results in a lower implied constant in the $\mathcal{O}\left(N^{-1/2}\right)$ convergence of the simulation error. An example of this type of approach is the information preserving DSMC (IP-DSMC) method proposed by Fan and Shen in [44], and developed by Sun and Boyd in [171, 172] for the low-speed non-equilibrium gas flows found in fluidic MEMS.

A second alternative is the hybrid DSMC method, which restricts the expensive particle simulation only to regions of the flow which are not in thermodynamic equilibrium. The hybrid DSMC method clearly yields the greatest computational savings when the non-equilibrium regions are as small as possible. Even when the non-equilibrium regions are not necessarily small, the hybrid DSMC method is still able to reduce the total simulation time when there are large density variations present in the flow. In these cases, the gas flow may be sufficiently dense in the equilibrium regions such that the cell size and time step restrictions on the DSMC method render the simulation intractable. The hybrid DSMC method is, therefore, well-suited to handle non-equilibrium flows with high density regions, which are common in hypersonic flows associated with re-entry vehicles. Recently, Schwartzentruber and Boyd in [156, 157] have demonstrated that the hybrid DSMC method does, in fact, achieve an appreciable cost savings over DSMC for certain high-speed, non-equilibrium gas flows.

A third alternative is the time relaxed Monte Carlo (TRMC) methods proposed by Pareschi and Russo [139, 140]. The key feature of the TRMC method is the novel time discretization of the collision operator in the Boltzmann equation, which approximates the higher order terms by an equilibrium velocity distribution. During the collision update step in the TRMC method, some of the simulated particles

are sampled directly from a local Maxwellian distribution for their post-collision velocities. As a consequence, the TRMC method allows time steps that are much larger than possible with DSMC. Further, the collision process preserves the correct asymptotic behavior in the limit as the time step tends toward infinity.

These benefits suggest that the TRMC method can offer appreciable reduction in the computational cost when compared to traditional DSMC methods for non-equilibrium gas flows in the slip regime ($Kn < 0.01$). Specifically, the collision timescale is typically many times smaller than the timescales associated with the macroscopic flow properties in the slip regime. Since the time step in traditional DSMC methods is limited by the collision timescale, many time steps are required to capture the changes in the macroscopic behavior of the gas. The TRMC method, in contrast, is able to use a much larger time step while remaining a physically consistent approximation to the collision process, which accounts for its cost-savings potential over traditional DSMC. To establish the accuracy of the TRMC method, Russo *et. al.* compares the new method to a proven DSMC method for the simulation of a spatially homogeneous gas [151] and a high-speed Couette flow [152].

Instead of improving the efficiency of the DSMC method, a fourth alternative is to use the so-called *extended fluid dynamic* approaches (*e.g.* the Burnett equations [22, 21]). The extended fluid dynamic approaches are essentially higher-order formulations to the Navier-Stokes equations that are able to extend the accuracy of the continuum-based methods to more rarefied flows in the transition regime. These higher-order formulations also involve higher-order derivatives in the governing partial-differential equations, which are more difficult to accurately simulate in practice. Hittinger [65] developed a novel scheme to simulate the extended fluid dynamic approaches using a system of hyperbolic-relaxation equations based on a closed

system of moment equations derived from kinetic theory [55, 94, 95]. Further development of the hyperbolic-relaxation scheme is currently being performed by Suzuki and van Leer [173]. The system of hyperbolic-relaxation equations only contain first-order derivatives and thus offer the following advantages over the higher-order partial-differential equations: (i) less restriction on the time step size for explicit schemes when the diffusion is numerically stiff; (ii) less sensitivity to the smoothness of the computational grid; and (iii) less communication between the computational cells (smaller stencil) making the scheme more efficient to implement on parallel computing architectures.

1.4 Outline

This investigation is organized in two parts corresponding to the two new approaches considered for the simulation of low-speed micro-scale gas flows. The first part, found in Chapter II, develops and tests the empirical corrections to Navier-Stokes simulation for the entire transition regime. The second part, found in Chapters III-VI, is devoted to the development of the quasi-Monte Carlo (QMC) particle method. In Chapter III, the basic theory concerning the convergence of the QMC method in general is reviewed. Chapter IV introduces a new construction of the low-discrepancy Weyl-Richtmyer sequence that is expected to offer some advantages when implemented in a QMC particle simulation. Further, the structure and computational cost of the basic algorithms used to generate the common low-discrepancy sequences needed in the QMC methods is also reviewed in Chapter IV. Given the difficulty associated with the design of a QMC particle method for general non-equilibrium gas flows, the method is only developed and tested in this investigation for free molecular, or collision-less flows. Specifically, the governing equations for free

molecular duct flows are presented in Chapter V, along with several different simulation techniques. These include: (i) the Markov chain simulation; (ii) the finite-state linear system solution; (iii) the Nyström method using Gauss-Legendre quadrature; (iv) the traditional test particle Monte Carlo method; and (v) the new QMC particle simulation. The most important result of Chapter V is that the QMC particle simulation proposed here is shown to achieve a near-linear error convergence, with significantly greater accuracy than the test particle Monte Carlo method. To determine the range of applicability of the method, the new QMC particle simulation is then tested in Chapter VI for 20 different duct geometries with a length to height ratio $0.5 \leq L \leq 10$. Although the convergence rate of the QMC particle simulation is found to decrease as the free molecular duct becomes narrower, the QMC particle simulation still demonstrates a faster convergence rate than the Monte Carlo methods. The cause of the performance loss in the QMC particle simulation is also considered, along with a possible correction, in Chapter VI. Finally, a summary of the major results of this investigation and brief outline of future research directions for the QMC particle simulation are given in Chapter VII.

Throughout the course of this research investigation, several original contributions are made by the author in an effort to improve the simulation of microscale gas flow. These new ideas and results appearing in the thesis include the following:

- (Section 2.4) A new technique is proposed for the construction of empirical models designed to correct the Navier-Stokes solution in the transition regime ($0.01 \leq Kn \leq 10$).

- (Section 2.5) New empirical models are found for Couette and Poiseuille flows using this construction technique.

- (Section 2.6) The new empirical models are tested much more thoroughly than previous models found in literature to better assess their predictive capabilities in the transition regime.

- (Section 4.1) A new implementation of the low-discrepancy Weyl-Richtmyer sequence, termed the BCF-3 sequence, is introduced.

- (Section 4.2) A process for selecting the set of irrational numbers required to generate the new BCF-3 sequence is presented.

- (Section 4.3) The BCF-3 sequence is shown to provide a noticeable improvement over the other Weyl-Richtmyer sequences found in the literature, when used in a QMC particle simulation.

- (Section 5.5) An original QMC particle simulation is developed to calculate the conductance probability of free molecular flow in a two dimensional duct.

- (Sections 6.1) The QMC particle simulation achieves a near-linear error convergence rate for a duct length to height ratio of two, irrespective of the choice for the low-discrepancy sequence of the method.

- (Sections 6.3) The QMC particle simulation is performed for a large number of duct geometries to best characterize the impact of the duct length on the performance of the method (in contrast with other applications found in the QMC literature that often present only a single test case).

- (Section 6.4) A new measure is proposed to quantify the extent of the non-physical correlation present between the dimensions of a low-discrepancy sequence, and to provide the minimum sequence length necessary for these correlation effects to be considered negligible.

- (Section 6.5) A hybrid quasi-Monte Carlo/Monte Carlo (QMC/MC) method is developed and subsequently shown to reduce the computational cost of the QMC particle simulation by a factor of 2 to 4.5.

CHAPTER II

EMPIRCAL CORRECTIONS TO THE NAVIER-STOKES SIMULATION

The computational challenge of simulating micro-scale gas flows has spawned many possible solution strategies. One of the most popular strategies involves correcting the continuum solution to include rarefaction effects. The corrections to the continuum solution can use alternative boundary conditions and transport closures, but do not usually affect the overall numerical solution technique. Thus, simulation of micro-scale gas flows can enjoy the same computational advantage as the underlying continuum method. Continuum methods based on the Euler and Navier-Stokes equations enjoy a rich and well-developed numerical simulation heritage [63, 174]. Without regard for accuracy, it is widely accepted that for a given flow and geometry the continuum simulation will reach its "solution" much faster than a physically accurate non-equilibrium method such as DSMC.

In the near continuum limit $Kn \to 0$, perturbation analysis demonstrates the need to relax the no-slip continuum boundary condition to allow for the presence of slip flow at the wall [109]. Therefore, in the limit of vanishingly small Kn, it is a physically and mathematical valid simulation technique to correct the continuum solution with a slip boundary condition. This correction to the traditional no-slip

boundary condition is referred to as a slip model. As the Knudsen number increases, the flow deviates further from local thermodynamic equilibrium and the continuum approximation in the transport closures for mass, momentum and energy break down. For the Poiseuille flows in this investigation, the error in the momentum transport closure is at least 25% when $Kn \geq 0.2$. Consequently, there are some continuum-based simulations which adopt empirical corrections to the transport closures in order to extend their applicable range to higher Knudsen numbers.

Corrected continuum methods that claim a range of applicability outside the near-continuum limit are not physically valid for large Knudsen numbers. However, the continuum methods may still give a reasonably accurate solution to the non-equilibrium problem. To a MEMS designer or fabricator, the path to the gas dynamic solution is irrelevant provided the final results are sufficiently accurate. If the corrected continuum method is accurate but non-physical, the degree to which the method is truly predicting the non-equilibrium phenomenon is highly dubious. It is especially true when the continuum corrections are derived from known non-equilibrium solutions provided *a priori*. This is most evident in the new unified models which combine empirical slip and transport corrections to obtain accurate Navier-Stokes solutions well outside the near continuum limit. The creators of such unified models claim that their continuum method can "predict" gas flow properties for all degrees of rarefaction $0 \leq Kn < \infty$ from continuum to free molecular flow.

Despite the philosophical objections to using a continuum solution in the free molecular regime, their extremely low computational cost makes them an inviting prospect. The continuum methods are so fast compared to physically accurate non-equilibrium techniques (*e.g.* DSMC and linearized Boltzmann), that there is a great deal of interest within the MEMS community to understand their range of applica-

bility. Sandia National Laboratories[1] is a leader in the field of MEMS design and fabrication. With their interest in MEMS, they supported this author during his summer practicum investigation of continuum corrections for micro-scale gas flows. The purpose of the summer investigation was to evaluate the accuracy of the new unified models that claim to extend the Navier-Stokes solution to the free molecular regime [111]. The initial study expanded to develop a new unified model with the goal of understanding its construction and assessing its ability to actually predict micro-scale gas flows [112].

The remainder of the chapter is devoted to continuing the investigation of unified models to correct continuum flows. Two new unified models are developed for Couette and Poiseuille flows. Similar to [112], these models are not developed as the "latest and greatest" replacements to those found in literature. Instead, they are developed only to understand the construction, accuracy, sensitivity and range of applicability. The organization of the chapter is as follows. In Section 2.1, the background of the continuum corrections is discussed including their history and potential error sources. In Section 2.2, the two most prominent unified models found in the literature for Couette and Poiseuille flows are given. In Section 2.3, an overview of the investigation procedure is provided. In Section 2.4, the procedure to determine the optimum Navier-Stokes solution that matches a known non-equilibrium result is explained. In Section 2.5, the construction of new unified models is detailed. In Section 2.6, the performance of the new unified models is analyzed for a variety of predictive cases. Finally in Section 2.7, the main points of the unified model investigation are summarized and recommendations for appropriate usage are given.

[1]Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

2.1 Background

The first continuum correction for non-equilibrium flows can be traced to Maxwell's founding work in gas kinetic theory [109]. Maxwell considered the behavior of gas molecules located within one mean free path of the wall undergoing idealized collisions with the surface. Maxwell's analysis, sometimes referred to as the mean free path method [54], demonstrates the existence of a finite slip velocity at the wall surface. In [109], the slip velocity u_s of the gas relative to the wall is determined to be

$$u_s = \frac{2-f}{f} \cdot \frac{2}{3}\ell_p \left.\frac{\partial u}{\partial n}\right|_{wall}, \tag{2.1}$$

where f is the fraction of particles undergoing a diffuse reflection with the wall surface, ℓ_p is the mean free path of the gas molecule, and $\left.\frac{\partial u}{\partial n}\right|_{wall}$ is the tangential velocity gradient in the direction normal to the wall. Maxwell originally treated the boundary condition as incalculable; but under the urging of a referee of his paper submitted to the Royal Society, he developed (2.1). Maxwell assumes two idealized types of wall collisions to model the boundary interaction, specular and diffuse. A specular reflection is similar to a ray of light reflecting off a mirror. The only change to the incident particle is that the normal component of the velocity reverses sign. The other velocity components are unaffected, thus there is no transfer of tangential momentum from the molecule to the wall. By contrast, a diffuse reflection is one that transfers the entire tangential momentum of the gas molecule to the wall. In a diffuse reflection, the incident molecule is considered to be absorbed by the wall for a period of time sufficient for the molecule to reach equilibrium with the wall before being re-emitted. As a consequence, the re-emitted molecule loses all memory of its previous trajectory. The reflected diffuse trajectory is thus equivalent to the

kinetic effusion of the gas from a reservoir at equilibrium with the wall [54]. Under these circumstances the gas molecule is said to be fully "accommodated" to the wall environment.

The boundary condition (2.1) is referred to as Maxwell's slip model. It demonstrates that the no-slip boundary condition used in the continuum solutions to the Navier-Stokes equations are not valid when the mean free path ℓ_p becomes significant. When the Navier-Stokes equations are corrected to include slip via (2.1), the results are accurate in the near continuum limit. The presence of slip velocity at the wall was demonstrated experimentally by Knudsen, and a review of his capillary tube measurements appears in [73]. Millikan noted the presence of slip velocity around the oil droplets in his famous electrostatic experiment [113]. Furthermore, in [113] Millikan conducts some of the earliest measurements of the empirical coefficient f used to model the gas-surface interaction in (2.1). Additional experimental work has prompted researchers to differentiate between the wall accommodation of an incident molecule's tangential momentum and thermal energy [154]. This allows an additional degree of freedom in approximating the gas-surface interaction and leads to a refinement of the form of Maxwell's slip model

$$u_s = \frac{2-\sigma_v}{\sigma_v} \cdot \frac{2}{3}\ell_p \left.\frac{\partial u}{\partial n}\right|_{wall},$$

where σ_v is the tangential momentum accommodation coefficient (TMAC). The TMAC has a similar role to original accommodation factor f used by Maxwell, except that the TMAC only measures the fraction of incident molecules whose tangential momentum is fully accommodated to the wall conditions. The energy accommodation is not needed in the boundary conditions for isothermal flow.

Modern research has continued improving Maxwell's slip model as interest in

rarefied gas flows has grown. The development of high altitude rocketry in the 1950's and 1960's, and the recent research in micro-flows for MEMS are responsible for the bulk of the modern slip model work. More accurate slip models have been developed, but they still rely on Maxwell's original framework

$$u_s = \frac{2-\sigma_v}{\sigma_v}\left(C_1 Kn \left.\frac{\partial u}{\partial \eta}\right|_{wall} + \frac{1}{2} C_2 Kn^2 \left.\frac{\partial^2 u}{\partial \eta^2}\right|_{wall}\right), \tag{2.2}$$

where C_1 and C_2 are coefficients that can be determined by any combination of analysis, numerical simulation, and experimental results. The Knudsen number $Kn = \ell_p/L$ appears in (2.2) because the wall normal component is normalized $\eta = n/L$ by the characteristic length scale L of the flow. The slip model (2.2) is considered a second-order approximation when an appropriate non-zero value for C_2 is selected. The coefficients C_1 and C_2 can be found analytically by a wide variety of approximations to the non-equilibrium, near wall solution [188, 39, 80, 25, 103]. The coefficients are also determined numerically using DSMC [138], the direct Boltzmann simulation [159] and the linearized Boltzmann simulation [133, 165]. Experimental results have been obtained for slip flows through long circular tubes [170]. A survey of the coefficient range is found to be $1.0 \leq C_1 \leq 1.1466$ and $-1.3089 \leq C_2 \leq 0.5$, as reported in [69].

Recent studies show that Maxwell's slip boundary condition breaks down around $Kn = 0.15$ [142]. Moreover, Piekos in [142] notes that there is an additional failure of the transport closure used in the Navier-Stokes equation. For example, the Poiseuille flow results in this chapter have at least a 25% error in the momentum transport when a slip model alone is used for $Kn \geq 0.2$. A group of researchers led by Karniadakis and Beskok have proposed unified models to provide not only slip boundary corrections but also momentum transport corrections [69, 15, 12, 11]. The

unified model allows for continuum corrections to be applied to increasingly rarefied flows with no loss of accuracy in certain special cases. In these unified models, the modern slip model (2.2) is further refined to allow for a slip coefficient C_s to depend on the Knudsen number

$$u_s = \frac{2-\sigma_v}{\sigma_v} C_s(Kn) Kn \left.\frac{\partial u}{\partial \eta}\right|_{wall}. \tag{2.3}$$

In addition, the apparent fluid viscosity μ' is also given a Knudsen number dependence

$$\mu' = \mu_0 C_\mu(Kn), \tag{2.4}$$

where μ_0 is the continuum viscosity of the fluid and $C_\mu(Kn)$ is a correction factor designed to recover the non-equilibrium momentum transport. The Knudsen number dependence of C_s and C_μ is empirical and is found by matching the Navier-Stokes solution to known non-equilibrium results. The continuum corrections provided by the unified models are reported to produce accurate Navier-Stokes solutions for $0 \leq Kn \leq 12$ in [12] and $0 \leq Kn < \infty$ in [15]. The computational time needed for standard non-equilibrium methods (DSMC and direct Boltzmann) is much greater than that to solve the corrected Navier-Stokes equations. Therefore it is of great interest to this investigation to understand and evaluate the accuracy of the unified models' continuum corrections.

2.1.1 Limitations to the slip model

The slip model is useful for understanding general trends of gas flows in the slip regime: increased mass flux through ducts, thermal creep, and decreased heat flux through the walls. However, its utility for calculating absolute (dimensional) flow properties is less clear. Some of the limits of using the slip model are discussed below

and should be kept in mind when considering the reported accuracy of the unified models.

Mean free path. The mean free path as an exact collision length scale is not well defined. From the original derivation of Maxwell's slip model (2.1) to the modern version (2.2) the 2/3 factor in the boundary condition has been unceremoniously dropped. The 2/3 factor appears in (2.1) because the average distance traveled in the direction normal to the wall by a reflected gas molecule is only 2/3 of a mean free path [109, 54]. Alternative collision length scales, similar to the boundary layer definition, can also be used to construct a slip model. For example, a collision length could represent the distance that 90% or 99% of the reflected molecules travel before colliding with the bulk flow. Regardless of the collision length scale chosen, a model can be constructed using Maxwell's reasoning with the sole difference being the leading coefficient C_1 in (2.2). This lack of certainty leads to some speculation on the true value C_1 by current slip model researchers.

Second-order corrections. The second-order term in (2.2) is designed to improve the accuracy of the slip coefficient and extend corrected continuum solution into more rarefied regimes. The second-order accuracy is only achieved in the limit of a small Knudsen number where the perturbation analysis is valid. Including the second-order terms offers no additional mathematical validity when $Kn \geq 1$ because the geometric series used in the perturbation analysis diverges. Furthermore, at any Knudsen number $Kn > 0$, the second-order slip models can not add any accuracy to the Navier-Stokes solution which is fundamentally equivalent to a first-order Chapman-Enskog expansion in Kn [28]. Adding a $O(Kn^2)$ correction is of no value given that the other second order terms are neglected when formulating the Navier-Stokes equation.

Gas	σ_v	Reference
He	~ 1	[4]
Ar	0.80 ± 0.01	[6]
	0.75–0.85	[5]
N_2	0.88 ± 0.01	[6]
	0.75–0.85	[5]
CO_2	0.75–0.85	[5]

Table 2.1: TMAC values reported by Arkillic *et. al.* for various gas species in micro-machined silicon channels.

Tangential Momentum Accommodation Coefficient. The exact value of the TMAC is difficult to measure because it is sensitive to so many factors, such as: surface roughness, environmental contamination, adsorbed layer composition, and surface age. A 5% error in the value of TMAC is equivalent to a 10% error in the slip coefficient C_1 in (2.2). Millikan measured gas-solid and gas-liquid surface accommodation coefficients with a range of $0.79 \leq \sigma_v \leq 1.00$ [113]. The minimum accommodation (79% diffuse reflection) is found for air interacting with a fresh shellac surface. The maximum accommodation (100% diffuse reflection) occurred for air interacing with a machined brass surface and several days old shellac. Arkillic *et. al.* measured the TMAC and established a range of values in several experiments with micro-machined silicon channels and various working fluids, reported in Table 2.1 The experimental methods of Millikan and Arkillic do not directly measure the TMAC. Instead, the TMAC is inferred from either the oil droplet velocity (Millikan) or the accumulated mass flow (Arkillic) when calculated with the continuum flow solution corrected with Maxwell's slip model boundary conditions (2.1). It is difficult to discern the accuracy of the slip model and the TMAC when their effects are lumped together in the same indirect measurement.

Direct measurements of the TMAC are made studying the surface reflection of high speed molecular beams in a vacuum environment [49]. Seidl and Steinheil

measure the TMAC for mono-energetic helium molecular beams reflected off various surfaces: single crystal copper (100), shellac, tungsten, gold, glass and sapphire [158]. The TMAC in [158] is found for the materials under common surface treatments and ranges in value from 1.16 for single crystal copper (100) with 5 micron grinding grooves to 0.67 for the same single crystal copper electrolytically polished. Seidl and Steinheil focus their study on the effect of the adsorbed layer on the solid surface and find similar concentrations of water and hydrocarbons for all materials tested. They proceed to remove the contaminates in the adsorbed layer by successive ion bombardment and annealing treatments which reduces the TMAC for the single crystal copper to 0.47. In order to fully remove the effect of the adsorbed layer in [158], an epitaxial layer of single crystal gold is grown on the copper surface and then annealed. The resulting treatment yields an extremely smooth surface on the atomic scale with a TMAC as low as 0.2. Lord conducts similar molecular beam experiments except with improved environmental conditions [99]. In [99], he obtains a TMAC of 0.2 for helium gas reflecting on a polycrystalline molybdenum surface without the use of an epitaxial layer [99]. Lord notes that the value of TMAC increases with the molecular weight of the gas species, and reports the following TMAC ranges for molybdenum and tantalum: He (0.20–0.46), Ne (0.31–0.59), Air (0.67–0.78), Kr (0.85) and Xe (0.95). The molecular beam experiments directly measure the TMAC and attempt to quantify the effects of complicated environmental influence on the gas-surface interaction. However, it is difficult to apply the molecular beam values of xenon gas on molybdenum to a micro-machined silicon channel, given the uncertain relationship between the experimental operating conditions and a typical MEMS device.

Gas-surface interaction. All slip models are based on Maxwell's initial assumption

[109] that each incident gas molecule can be classified as either having zero accommodation (specular reflection) or full accommodation (diffuse reflection) regardless of incident angle. The molecular beam experiments mentioned in the previous section [158, 99] show a dependence of the measured TMAC value on the incident angle. In some cases, the TMAC varies by 30% or more for incident angles ranging from 10 degrees to 70 degrees from the surface normal. The more glancing collisions (high incident angle relative to the surface normal) yield more specular reflections, thus lowering the observed TMAC. Cercignani and Lampis propose a phenomenological model that includes more surface physics than Maxwell's original assumption [26]. Lord includes even more physics to the model in [26] to produce what is often referred to as the CLL model [100, 101]. The CLL model is well-suited for use in non-equilibrium calculations like DSMC; however, it can not be incorporated in a continuum method because of predicate assumption of a near-equilibrium velocity distribution function. Yamanishi *et. al.* have also proposed more complicated gas-surface boundary conditions using a database of interactions simulated with a direct molecular dynamics method [191]. The evidence of more complicated gas-surface interaction than approximated by Maxwell's original slip model should prompt the user of such continuum corrections to be wary of reporting any results to very high accuracy.

2.2 Unified Models

All gas flows, in equilibrium or not, must satisfy the conservation laws of mass, momentum, and energy. For continuum flows, these conservation laws are represented by a set of 5 differential equations, with additional transport closures for momentum and energy and a state equation, for which there is no known general

solution. Assuming isothermal flow and the Newtonian shear stress closure, the conservation of energy is automatically satisfied, which results in a system of equations consisting of the continuity equation (2.5), the Navier-Stokes equation (2.6) and the shear stress closure (2.7).

$$\frac{\partial \rho}{\partial t} + \frac{\partial \rho v_i}{\partial x_i} = 0, \tag{2.5}$$

$$\frac{\partial v_i}{\partial t} + v_j \frac{\partial v_i}{\partial x_j} = \frac{1}{\rho}\left(\frac{\partial \tau_{ij}}{\partial x_j} - \frac{\partial p}{\partial x_i}\right) + f_i, \quad \text{where} \tag{2.6}$$

$$\tau_{ij} = \mu_0 \left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\right). \tag{2.7}$$

In the above equations, ρ is the fluid density, v is the fluid velocity vector, p is the fluid pressure, f is the acceleration due to an external body force acting on the fluid, μ_0 is the fluid viscosity, and τ_{ij} is the stress tensor.

The unified models developed recently use continuum corrections to the Navier-Stokes solution to approximate non-equilibrium flows in the slip to free molecular range $Kn > 0.001$ [69, 15, 12, 11]. The idea is not new as researchers have developed approximations to Couette flows for all degrees of rarefaction by solving the low-order moments of the Boltzmann equation assuming a model velocity distribution function [92]. For certain flow geometries, the relative shape of the macroscopic velocity profile does not significantly change with Knudsen number. This is especially true for the canonical viscous geometries of Poiseuille and Couette flows. Poiseuille and Couette flows permit analytical solutions to the continuum fluid equations (2.5), (2.6) and (2.7) under the assumptions of constant density and steady flow for a two dimensional duct with constant area. The continuity equation (2.5) is no longer needed and the Navier-Stokes equation and shear stress closure combine to yield a single ordinary differential equation for each case. The continuum solution to body force driven

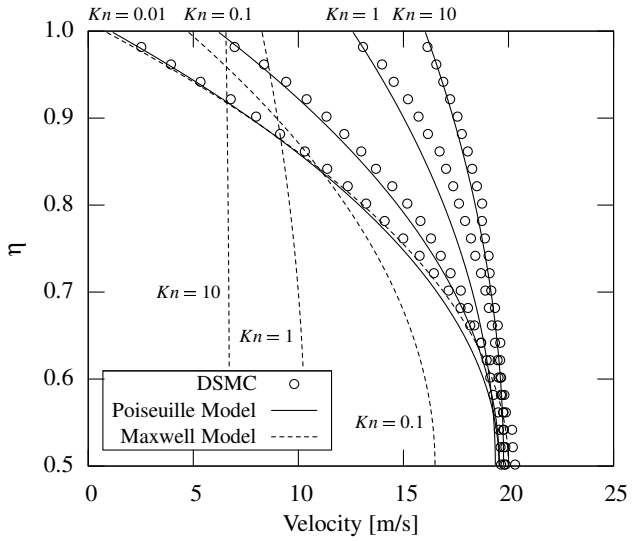

Figure 2.1: Poiseuille velocity profiles for $0.01 \leq Kn \leq 10$.

Poiseuille flow satisfies

$$\frac{\partial^2 u}{\partial \eta} = -\frac{\rho f h^2}{\mu} \quad \text{with} \quad u(0) = u(1) = 0, \tag{2.8}$$

where u is the tangential velocity in the duct, and h is the duct height used to normalize the wall normal direction $\eta = y/h$. Similarly, the continuum solution to Couette flow satisfies

$$\frac{\partial^2 u}{\partial \eta} = -\frac{\rho f h^2}{\mu} \quad \text{with} \quad u(0) = U_0 \text{ and } u(1) = U_1, \tag{2.9}$$

where U_1 and U_2 are the lower and upper velocities of the wall boundaries. In Figure 2.1, Poiseuille velocity profiles are plotted for a non-equilibrium solution (DSMC - circles) and the best possible continuum corrections to the Navier-Stokes equation (Poiseuille Model - solid line). For all Knudsen numbers $0.01 \leq Kn \leq 10$, the non-equilibrium velocity profiles are nearly parabolic. Thus, the Navier-Stokes solution to Poiseuille flow (2.8), which is strictly a symmetric parabola, can be corrected to yield a close fit to the non-equilibrium results. Similarly in Figure 2.2, Couette profiles are plotted using the same simulation techniques and conditions. Now for all

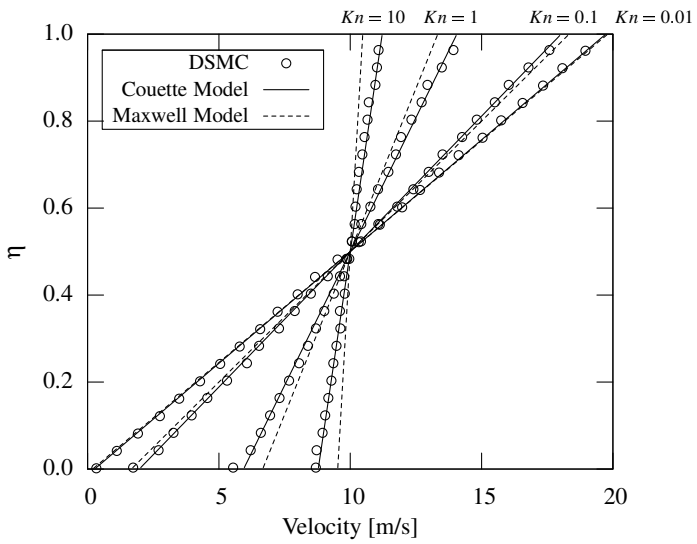

Figure 2.2: Couette velocity profiles for $0.01 \leq Kn \leq 10$.

Knudsen numbers $0.01 \leq Kn \leq 10$, the non-equilibrium velocity profiles are nearly linear. Thus, the Navier-Stokes solution to Couette flow (2.9), which is strictly linear, can also be corrected to yield a close approximation to the non-equilibrium results. The non-equilibrium profiles in Figures 2.1 and 2.2 show an extra curvature in the near wall region, especially for the transitional flow cases $Kn = 0.1$ and $Kn = 1$. This effect is termed the Knudsen layer and is limited to the region within a few mean free paths of the wall [73, 154, 179]. The Knudsen layer is a consequence of the non-equilibrium relaxation of the reflected gas molecules from the wall to the streaming bulk flow conditions.

All the non-equilibrium Poiseuille flows in Figure 2.1 exhibit a non-zero slip velocity at the wall that increases with Knudsen number. The goal of the slip model boundary condition (2.3) is to correct the continuum Navier-Stokes solution to capture this non-continuum effect. The unified models assume the slip velocity can be corrected for all degrees of flow rarefaction if one knows the non-equilibrium solution *a priori*. Replacing the no-slip boundary condition in (2.8) with the slip boundary

condition (2.3) allows one freedom to choose $C_s(Kn)$ to match every known non-equilibrium slip velocity. This is the crux of the construction of the unified models.

It is important to note, that the Navier-Stokes solution can not be tuned to match the non-equilibrium results $Kn \geq 0.2$ via the slip model alone. As Maxwell's slip model (2.1) demonstrates in Figure 2.1, the curvature approximation of the Navier-Stokes velocity profile worsens as the Knudsen number increases. The curvature of the Navier-Stokes solution is set by the flow constants on the right hand side of (2.8). Only the fluid viscosity is not explicitly used in a non-equilibrium calculation because it is only a by-product of the continuum shear stress closure. For this reason, the unified models must include a correction to the viscosity (2.4) in order for the continuum solutions to capture the non-equilibrium curvature when the transport closure breaks down. Therefore, the freedom to choose the values of the two empirical coefficients $C_s(Kn)$ and $C_\mu(Kn)$ allows for a corrected Navier-Stokes solution to accurately approximate almost any non-equilibrium Poiseuille result.

Similarly, all the non-equilibrium Couette flows in Figure 2.2 exhibit a non-zero slip velocity at the wall that increases with Knudsen number. The continuum Couette velocity profile obtained by solving the linear homogeneous differential equation in (2.9), is simply a straight line matching the two wall speeds. The addition of the slip model (2.3) does not change the character of the solution. The only permissible solutions of the corrected continuum solution are linear; however, the freedom to choose $C_s(Kn)$ allows for any slope to be matched. As Figure 2.2 illustrates, the ability of the continuum correction to match any slope is sufficient to ensure accurate approximation to any non-equilibrium Couette result. If the ideal slope is selected for the Navier-Stokes solution to match the non-equilibrium result, the average shear stress in the channel is still incorrect due to the failure of the continuum transport

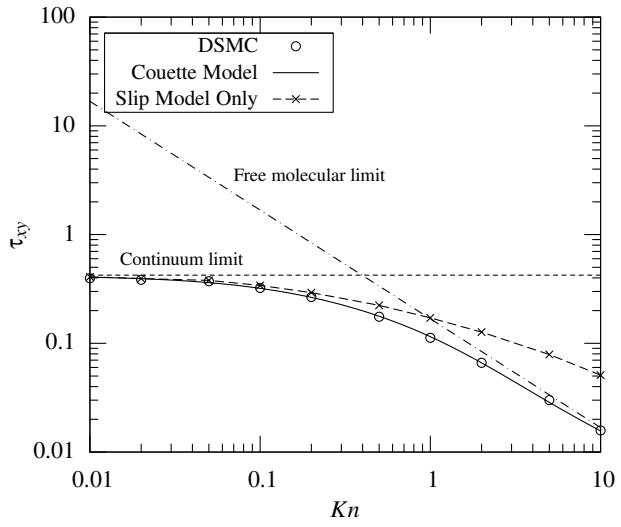

Figure 2.3: Comparison of the continuum shear stress predicted by the unified model and by slip model alone.

closure, see Figure 2.3. When the Navier-Stokes solution is only corrected by a slip model, the shear stress is 50% greater than the non-equilibrium result at $Kn = 1$, and 3 times greater at $Kn = 10$. Therefore, it is necessary for the Navier-Stokes solution to use the viscosity correction (2.4) to overcome the error in the transport closure. For both Poiseuille and Couette flows, the non-equilibrium solutions maintain a similar enough shape in the range $0.01 \leq Kn \leq 10$ for accurate Navier-Stokes solutions to exist. However, accurate corrections to the Navier-Stokes solution at $Kn \geq 0.2$ can only be obtained if both the slip coefficient C_s and the viscosity correction C_μ are used.

2.2.1 Unified Poiseuille model

Karniadakis and Beskok propose a unified model for Poiseuille flow involving empirical corrections (2.3) and (2.4) to the Navier-Stokes equation [69, 15]. The slip

coefficient C_s is modeled by

$$C_s(Kn) = \frac{1}{1 - b_0 Kn}, \tag{2.10}$$

where b_0 is a free parameter selected to be $b_0 = -1$ for Poiseuille flow. The viscosity correction is modeled by

$$C_\mu(Kn) = \frac{1}{1 + \alpha Kn}, \tag{2.11}$$

where α is an empirically determined rarefaction parameter. This model will be referred to as the KB model for the remainder of the investigation. The value $b_0 = -1$ in (2.10) is found by fitting the normalized Navier-Stokes velocity profile to linearized Boltzmann results. Note that normalizing the velocity profile isolates the slip correction from the viscosity correction. Specifically choosing $b_0 = -1$ has an additional benefit in that it is a second-order correction to Maxwell's slip model in the near continuum limit $Kn \to 0$. Other asymptotic limits can be satisfied with the selection of α. An additional model can be constructed to ensure α yields the correct results in the limits $Kn \to 0$ and $Kn \to \infty$ [15]

$$\alpha = \alpha_0 \frac{2}{\pi} \tan^{-1}(\alpha_1 Kn^\beta). \tag{2.12}$$

Here α_0 is the free molecular value and α_1 and β can be selected to match any given non-equilibrium solution in the transition regime. Alternatively, the model (2.12) can be forsaken altogether by selecting α directly from a database of linearized Boltzmann solutions [12].

2.2.2 Unified Couette model

Similarly, Bahukudumbi, Park and Beskok propose a unified model for Couette flow, referred to as the BPB model for the remainder of the investigation [12, 11]. The aesthetic and second order accuracy found in the KB model is dropped in favor

of a more direct modeling strategy. The Knudsen number dependence of the slip coefficient adopts a four parameter arctangent form

$$C_s(Kn) = a_1 + a_2 \tan^{-1}(a_3 Kn^{a_4}), \tag{2.13}$$

with the coefficients determined empirically to be $a_1 = 1.2977$, $a_2 = 0.71851$, $a_3 = -1.17488$ and $a_4 = 0.58642$. The viscosity correction does not use the form of (2.4) explicitly, rather the corrected shear stress is obtained by the non-dimensional form

$$\Pi_{xy} = \frac{\tau_{xy}}{(\tau_{xy})_\infty} = -\frac{bKn^2 + 2cKn}{bKn^2 + dKn + c}, \tag{2.14}$$

where $(\tau_{xy})_\infty$ is the shear stress present in free molecular Couette flow and the coefficients are $b = 0.529690$, $c = 0.602985$ and $d = 1.627666$. The free molecular shear stress in determined analytically from gas kinetic theory

$$(\tau_{xy})_\infty = \rho(U_1 - U_0)\sqrt{\frac{kT}{2\pi m}}. \tag{2.15}$$

The advantage of the form in (2.14) is that it is straightforward to enforce the correct asymptotic limits on the shear stress.

2.3 Investigative Method

The KB model and the BPB model are designed to correct the Navier-Stokes solution for Poiseuille and Couette flows with any degree of rarefaction [69, 15, 12, 11]. However, the successful construction of both unified models requires the availability of known non-equilibrium solutions. This observation begs the immediate question: what good is an ultra-fast Navier-Stokes solution to a non-equilibrium flow if one needs the slower non-equilibrium solution first? The whole point is to avoid calculating the expensive non-equilibrium solution in the first place. While the unified models are not truly predicting non-equilibrium flows, they can offer computational savings

in some instances. If a geometry is fixed, but the scale of the problem is allowed to vary (*e.g.* the slider bearing design of the magnetic reader of a hard drive), the unified model could be useful interpolating flow parameters between non-equilibrium solutions. Another use for the unified models is for flow geometries that are reasonably close to the non-equilibrium solutions used in the models' construction, such as the oscillating Couette flow [12] and the slider bearing flow [11].

The KB model and the BPB model are reported to yield continuum solutions with great accuracy for any non-equilibrium conditions. In particular, [15] states the following:

> "... we have developed a simple physics-based *unified model* that predicts the velocity distribution, the volumetric and mass flow rates, as well as the pressure distribution in channel, pipes, and duct flows (of general aspect ratio) for the entire flow regime (*i.e.*, $0 \leq Kn < \infty$)."

This author believes the use of the word "predict" is not valid. The amount of effort spent generating the non-equilibrium solutions for the models' construction would seem to indicate that the unified models are not actually predicting non-equilibrium flows. Instead, the unified models are just carefully tuned to reproduce the non-equilibrium results. Bird issues a seemingly prescient warning in 1994 (p. 184 of [16]) to future aficionados of the unified models:

> "The fact that solutions are available for the two limiting cases of collisions and continuum flows means that superficially good results may be obtained from physically unreal methods that happen to provide a fortuitously good curve fit between these known limits. Particular scrutiny should be given to solutions that are based on approximations which introduce adjustable parameters."

By simulating with a unified model, one gives up the physical accuracy of a true non-equilibrium method in favor of computational speed of the Navier-Stokes solution. However to initially construct a unified model, one forgoes the computational speed of

the Navier-Stokes solution in favor of the physical accuracy of a true non-equilibrium method. The unified model essentially is caught in a rarefied gas dynamics perversion of *The Gift of the Magi* [61].

If it is possible to obtain an accurate non-equilibrium solution with continuum corrections to the Navier-Stokes equation, it would represent a tremendous advantage for evaluating fluidic MEMS designs. Given the need for non-equilibrium solutions *a priori*, the unified models do not seem to deliver that advantage in cases of true prediction. Despite the disheartening outlook for unified models, as indicated earlier, there are flows for which the unified models should provide an accurate level of prediction. Specifically, gas flows close in either geometry or operating conditions to the non-equilibrium solutions used to construct the unified model. It is therefore important to understand how these unified models are constructed, and establish the accuracy of deviations from the known solutions. In order to accomplish these goals, the following investigative procedure is proposed:

1. Generate a database of non-equilibrium Couette and Poiseuille flows.
2. Find the optimum model coefficients C_s^* and C_μ^* that produce the Navier-Stokes solution that best matches the non-equilibrium results.
3. Form new unified model laws that capture the functional dependence of the optimum coefficients found in Step 2.
4. Test the new unified models for various predictive cases.

2.3.1 Database construction

The database of non-equilibrium solutions used to generate the new models consists of one dimensional argon and nitrogen Couette and Poiseuille flows. The DSMC solutions used for the database are obtained from a modified version of the one di-

mensional code provided by Bird [16]. All flows in the database are low speed flows, the Poiseuille flows are driven to a maximum velocity of 20 m/sec and the Couette flows use a difference in wall velocities of 20 m/sec. Knudsen numbers of 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, and 10 are simulated by adjusting the operating density of the working fluid, while keeping the geometry and computational domain constant. The walls are simulated with a temperature of 273 K and a fully diffuse, gas-surface interaction yielding a TMAC of unity. The DSMC simulation is one dimensional which implies that the velocity distribution function is everywhere uniform along planes parallel to the walls. As a result, a pressure gradient cannot be used to drive the Poiseuille flow cases; instead, the accelerative body force f is used. The driving force varies with Knudsen number, and is found through trial and error until the DSMC results appear to have a maximum velocity of 20 m/sec. The collision dynamics are simulated using the variable soft sphere model with the collision parameters from Bird, and the rotational energy exchange for the nitrogen gas flows is simulated with the Larsen and Borgnakke model [16]. All DSMC simulations use 150 cells and 4500 simulated particles, except for the case of Couette flow at $Kn = 0.01$, which uses 300 cells and 9000 simulated particles. The maximum cell length is less than one third of a mean free path. The time step is chosen so that a particle will cross a cell in an average of three time steps. The results are sampled for 20 million time steps with the typical statistical scatter in the velocity profile less than 1%.

2.3.2 Optimum model coefficients

Given the freedom to choose any values for the two empirical corrections included in the unified model C_s (2.3) and C_μ (2.4), one can alter the Navier-Stokes solution to match almost any non-equilibrium Couette and Poiseuille result. For Couette

flow, where the non-equilibrium solution remains nearly linear, the line that best matches the DSMC data is found using the linear least squares technique. The slope of the best-fitting line corresponds uniquely to a specific value of C_s. The value of C_s that produces the best-fitting Navier-Stokes solution is termed the optimum slip coefficient C_s^*. Once the slope of the best fitting line is determined for Couette flow, the optimum viscosity correction C_μ^* is found such that the corrected Navier-Stokes solution (2.18) and the DSMC solution have the same channel-averaged shear stress.

Similarly for Poiseuille flow, where the non-equilibrium solution remains nearly parabolic, the channel symmetric parabola that best matches the DSMC data is found using the linear least squares technique. The two free parameters that define the best-fitting parabola correspond uniquely to specific values of C_s and C_μ. These two values for the unified model that produce the best-fitting Navier-Stokes solution are also termed the optimum slip coefficient C_s^* and C_μ^* for Poiseuille flow. Unlike Couette flow, the viscosity correction C_μ for Poiseuille flow is not needed to match the non-equilibrium wall shear stress. Both the continuum Navier-Stokes solution and the non-equilibrium DSMC simulation conserve momentum exactly throughout the domain. Therefore, the method of simulation for the momentum transport has no effect on the shear stress at the boundary surface. The wall shear stress at the boundary surface for a one dimensional, body force driven Poiseuille flow is determined entirely by the fluid density, body force and channel height (2.19). Corrections to the Navier-Stokes solution via the unified model coefficients C_s and C_μ can match the non-equilibrium velocity profiles to within an L_2 error norm of 2% for both Couette and Poiseuille flows in the range of $0.01 \leq Kn \leq 10$. The optimum coefficients C_s^* and C_μ^* are found for the unified model using the least squares technique in Section 2.4 for every non-equilibrium case in the DSMC database.

2.3.3 Unified model construction

After the optimum coefficients C_s^* and C_μ^* are determined for each case in the DSMC database, model laws are found to approximate the Knudsen number dependence of the coefficients. Several non-linear model laws are tested in an attempt to find the unified model that best matches the DSMC data for argon gas throughout the range of rarefaction tested ($0.01 \leq Kn \leq 10$). A new model construction technique is proposed in Section 2.5 using an importance weighting of the optimum coefficient data in conjunction with the Levenberg-Marquardt non-linear least squares minimization. The error sensitivity for deviations in the optimum coefficients is found for all Knudsen numbers. The first order estimation of the sensitivity is used as the importance weight when fitting the model to the optimum coefficients. The resulting non-linear models show marked improvement in uniform accuracy over all degrees of rarefaction compared to the previous models constructed by this author [112]. The new unified models are within 1% of the best possible L_2 error of the velocity profile when obtained directly from the optimum coefficients.

2.3.4 Predictive cases

Great care is taken to ensure the empirical Couette and Poiseuille models developed in this investigation for the slip and viscosity model coefficients capture the non-equilibrium flows in the DSMC database. However, the accuracy of the models for cases within the database is not a measure of the models' applicability, only of the data fit correlation. The new unified models are not predicting the non-equilibrium flows from the database, they are just carefully tuned to reproduce them. Since the new unified models are purely empirical and the continuum solutions based on them break down as the flow deviates from equilibrium, they should be suspect when

predicting flows outside the database at large Knudsen numbers. In order to assess the actual predictive power of the new unified Couette and Poiseuille models, five types of test cases outside the DSMC database are selected to illustrate different non-equilibrium challenges. The models developed in this investigation are used to predict these flows and then compared to DSMC results. The first cases involve interpolation and extrapolation of the database for both Couette and Poiseuille argon flows at Knudsen numbers of 0.7 and 20. Second, a combination Couette and Poiseuille flow is simulated for argon gas at $Kn = 1$. Third, the tangential momentum accommodation coefficient (TMAC) is changed from unity to 0.8 and 0.5 for both Couette and Poiseuille argon flows. Fourth, helium gas is used as the working fluid for the Couette and Poiseuille flows at $Kn = 1$. Helium has a molecular weight about one tenth of that of argon, so the resulting most probable molecular velocity is about three times that of argon. Finally, a body force driven flow with uniform suction and injection normal to the walls is simulated at $Kn = 1$. While the solution is still one dimensional, it is the only flow in this investigation that has a non-zero convective acceleration. Most multidimensional flows, or flows with complex geometry have a non-zero convective acceleration, so the ability of the new unified Poiseuille model to capture the physics change in this flow is an indication of the applicability of the new models toward more complex flows. The analytical solution to the body force driven flow with a uniform suction and injection velocity V_0 at the walls is given in the following equations [187]:

$$u_\chi = \frac{F'}{Re}\left[D\left(e^{\eta Re} + \Lambda Re - 1\right) + \eta + \Lambda\right], \quad \text{where}$$

$$D = \frac{1 + 2\Lambda}{1 - \Lambda Re - (1 + \Lambda Re)e^{Re}},$$

$Re = \rho V_0 h/\mu'$ is the non-dimensional Reynolds number based on the cross flow velocity V_0, $\Lambda = \frac{2-\sigma_v}{\sigma_v} KnC_s(Kn)$ is the combined non-dimensional slip coefficient, μ' is the apparent fluid viscosity, and $F' = \rho f h^2/\mu'$ is the force term.

2.4 Least squares fit

The Navier-Stokes equation can be solved analytically for most Couette and Poiseuille flows that are one dimensional, fully developed, steady, isothermal and constant density. These are the same conditions found in the DSMC simulations. Using the empirical slip model (2.3) and viscosity correction (2.4) introduced earlier, the resulting Navier-Stokes solutions to the ordinary differential equations in (2.8) and (2.9) for the flow velocity and shear stress are

$$u_c(\eta) = U_0 + \frac{U_1 - U_0}{1 + 2\Lambda} (\eta + \Lambda), \tag{2.16}$$

$$u_p(\eta) = -\frac{F'}{2} \left(\eta^2 - \eta - \Lambda\right), \tag{2.17}$$

$$(\tau_{xy})_c = \frac{\mu'}{1 + 2\Lambda} (U_1 - U_0)/h \quad \text{and} \tag{2.18}$$

$$(\tau_{xy})_p = -\rho f h (\eta - 1/2). \tag{2.19}$$

In the above equations, u_c and u_p are the Couette and Poiseuille velocity profiles, $(\tau_{xy})_c$ and $(\tau_{xy})_p$ are the Couette and Poiseuille flow shear stress, U_0 and U_1 are the lower and upper wall velocities for the Couette flow, $\eta = y/h$ is the wall normal coordinate non-dimensionalized by the channel height h, $\Lambda = \frac{2-\sigma_v}{\sigma_v} KnC_s(Kn)$ is the combined non-dimensional slip coefficient, and $F' = \rho f h^2/\mu'$ is the body force term. The purpose of the empirical coefficients is to tune the continuum-based Navier-Stokes solutions to yield an approximation to the non-equilibrium DSMC results in the database. The Navier-Stokes velocity profile for Couette flow is simply a straight line with the following constraint that the velocity at the midpoint of the channel

must be the average of the two wall velocities. Therefore, any line of the form in (2.20) is a valid Navier-Stokes solution for shear-driven flow

$$u_c(\eta) = G_c\left(\eta - \frac{1}{2}\right) + \frac{1}{2}(U_0 + U_1), \tag{2.20}$$

where G_c is a free parameter that characterizes the family of solutions for different slip coefficients.

We can select the Navier-Stokes solution from this family that best fits the DSMC data by performing a linear least squares fit of the DSMC velocity data. The linear least squares technique minimizes the L_2 error norm between the known data (DSMC) and the approximate linear model (Navier-Stokes) [176]. The L_2 error norm is the non-dimensional measure of a the velocity profile error

$$L_2 = \frac{1}{\bar{v}N}\sqrt{\sum_{i=1}^{N}\left[(v_{mc})_i - (v_{ns})_i\right]^2} \quad \text{where} \quad \bar{v} = \frac{1}{N}\sum_{i=1}^{N}(v_{mc})_i\,, \tag{2.21}$$

and $(v_{mc})_i$ and $(v_{ns})_i$ are the DSMC and Navier-Stokes velocity in cell i respectively. The Navier-Stokes solution with the lowest L_2 error is defined by

$$G_c = \frac{(S_{uy})_c}{(S_{yy})_c},$$

where

$$(S_{yy})_c = \sum_{i=1}^{N}\left(\eta_i - \frac{1}{2}\right)^2$$

and

$$(S_{uy})_c = \sum_{i=1}^{N}\left[(u_i)_c - \frac{1}{2}(U_1 + U_2)\right]\left(\eta_i - \frac{1}{2}\right).$$

In the above equations, η_i and $(u_i)_c$ are the non-dimensional position and velocity respectively in the i^{th} DSMC cell, and N is the total number of cells. Once the free parameter G_c is found for the best fitting solution, it corresponds uniquely to the slip

coefficient. The optimum slip coefficient necessary to capture the non-equilibrium DSMC profile in a least squares sense is determined by

$$C_s(Kn) = \frac{U_2 - U_1 - G_c}{2\left(\frac{2-\sigma_\nu}{\sigma_\nu}\right) KnG_c}.$$

As mentioned earlier, matching the velocity in Couette flow is only half the problem. In order to capture the correct shear stress, the viscosity correction C_μ is selected to reproduce the average shear stress of the DSMC data. The free parameter G_c characterizing the best fitting profile is used to determine the viscosity correction

$$C_\mu(Kn) = \frac{h}{\mu_0 N G_c} \sum_{i=1}^{N} (\tau_{xy})_i,$$

where $(\tau_{xy})_i$ is the shear stress in the i^{th} DSMC cell.

The Navier-Stokes solution for the Poiseuille flow velocity profile is a channel symmetric parabola. This means that there are two free parameters G_{p1} and G_{p2} to characterize the family of valid Navier-Stokes solutions, with different boundary conditions and transport closures

$$u_p(\eta) = G_{p2}(\eta^2 - \eta) + G_{p1}.$$

Similar to Couette flow, the values of G_{p1} and G_{p2} can be found that best match the non-equilibrium solution by performing a linear least squares fit to each DSMC Poiseuille flow case in the database:

$$\begin{bmatrix} G_{p1} \\ G_{p2} \end{bmatrix} = \begin{bmatrix} N & (S_y)_p \\ (S_y)_p & (S_{yy})_p \end{bmatrix}^{-1} \begin{bmatrix} (S_u)_p \\ (S_{uy})_p \end{bmatrix},$$

where

$$(S_y)_p = \sum_{i=1}^{N} \left(\eta_i^2 - \eta_i\right),$$

$$(S_{yy})_p = \sum_{i=1}^{N} \left(\eta_i^2 - \eta_i\right)^2,$$

	Couette Flow				Poiseuille Flow			
	C_s^*		C_μ^*		C_s^*		C_μ^*	
Kn	Ar	N_2	Ar	N_2	Ar	N_2	Ar	N_2
0.01	1.821	1.907	0.970	0.960	1.493	1.318	1.006	1.017
0.02	1.850	1.688	0.974	0.983	1.415	1.259	0.992	0.982
0.05	1.160	1.274	0.981	1.006	1.254	1.277	0.940	0.978
0.1	1.214	1.198	0.948	0.974	1.188	1.126	0.877	0.906
0.2	1.118	1.035	0.911	0.917	0.972	0.953	0.722	0.749
0.5	0.894	0.873	0.785	0.807	0.671	0.661	0.463	0.492
1	0.735	0.695	0.656	0.668	0.469	0.459	0.297	0.314
2	0.584	0.556	0.522	0.534	0.308	0.299	0.173	0.182
5	0.437	0.427	0.381	0.398	0.170	0.168	0.080	0.086
10	0.365	0.324	0.310	0.298	0.107	0.101	0.044	0.044

Table 2.2: Table of the optimum coefficients C_s^* and C_μ^* for non-equilibrium Couette and Poiseuille flows.

$$(S_u)_p = \sum_{i=1}^{N} (u_i)_p,$$

and

$$(S_{uy})_p = \sum_{i=1}^{N} (u_i)_p \left(\eta_i^2 - \eta_i\right).$$

Once the free parameters G_{p1} and G_{p2} are found, they uniquely determine the slip coefficient and viscosity model coefficient that will best capture the non-equilibrium velocity profile in a linear least squares sense:

$$C_s(Kn) = -\frac{G_{p1}}{G_{p2}\left(\frac{2-\sigma_\nu}{\sigma_\nu}\right) Kn},$$

and

$$C_\mu(Kn) = -\frac{\rho f h^2}{2 G_{p2} \mu_0}.$$

Using the database of DSMC cases as reference, it is possible to generate the best slip and viscosity model coefficients to match the Navier-Stokes solution to each non-equilibrium solution. The optimum coefficients are found to match all the non-equilibrium flows in the database to within an L_2 error of 2.5%.

2.5 New Model Laws

In the previous section, the optimum coefficients C_s^* and C_μ^* are found in order to fit the corrected Navier-Stokes solution to every case in the DSMC database and listed Table 2.2. These optimum coefficients are found using the standard linear least squares method and represent the Navier-Stokes solution that fits the non-equilibrium data with the minimum L_2 error. New unified models similar to the KB and BPB models can be constructed with the optimum coefficients C_s^* and C_μ^* found in the previous section. A unified model is designed to capture the Knudsen number dependence of the continuum corrections (2.3) and (2.4) to the Navier-Stokes solution. The goal is to have an explicit functional form for $C_s(Kn)$ and $C_\mu(Kn)$ that yields accurate Navier-Stokes approximation to all the non-equilibrium flows in the database. One possible model would recover the optimum coefficient $C(Kn^*) = C^*$ when at the same conditions Kn^* as the non-equilibrium database using a piece-wise approximation to the data (*e.g.* cubic spline). However, the exact match of the optimum coefficient C^* is not necessary given the nature of the approximation; more often a single functional form is preferred.

There are several advantages to selecting a single function model over a cubic spline or other piece-wise approximation. First, any scatter present in the optimum coefficients can potentially be smoothed out by a single function designed to minimize its distance to the data. A cubic spline must intersect every given data point, thus its representation of the scatter may yield unnecessary fluctuations in the resulting curve. Second, while both methods are suitable for interpolation of the known non-equilibrium solution, the single function model can be trained for extrapolation by including known asymptotic results. For example, the BPB model is designed to

recover the correct shear stress in both the continuum $Kn \to 0$ and free molecular $Kn \to \infty$ limits. Third, unlike a spline model, the single function can enforce other known physical properties of the process (*e.g.* monotonicity). Finally, the importance of each non-equilibrium case in the database may not be equal when trying to find a continuum correction with uniform accuracy across all degrees of rarefaction. The construction of the single function model enables the weighting of the relative importance of each non-equilibrium solution. No such weighting is possible with a piece-wise approximation, all the points are of equal weight.

2.5.1 Non-linear data fitting

The approximate functional dependence of the optimum coefficients $C_s(Kn)$ and $C_\mu(Kn)$ is found by a least squares minimization technique. The optimum coefficients C_s^* and C_μ^* in Table 2.2 demonstrate a monotonic decrease in magnitude with increased rarefaction Kn. This observation precludes the use of a polynomial model function as an accurate representation of the data. In fact, most model functions with linear parameters provide inadequate representations of the data in Table 2.2. Thus, non-linear models are selected to construct the new unified models. The drawback to non-linear models is that the simple linear least squares technique of Section 2.4 is no longer applicable. Instead, the iterative Levenberg-Marquardt method [93, 107] is used to minimize the least squares error of the non-linear model relative to the optimum coefficients C_s^* and C_μ^*.

The Levenberg-Marquardt method finds the best-fitting, non-linear model by combining two common minimization techniques. First, consider a general non-linear model for the optimum coefficients

$$C = C(Kn; \mathbf{a}),$$

where C is either unified model coefficient (C_s or C_μ) and $\mathbf{a} = (a_1, \ldots, a_m)$ are the unknown, non-linear model parameters. The quality of the fit of a non-linear model given a specific parameter set $\mathbf{a}$ is quantified by a merit function χ^2 defined by

$$\chi^2(\mathbf{a}) = \sum_{i=1}^{N} \left[\frac{C^* - C(Kn; \mathbf{a})}{\sigma_i} \right]^2, \tag{2.22}$$

where the model $C(Kn; \mathbf{a})$ fits the N data points from the DSMC database, and σ_i is an importance weight of the i^{th} datum. The best fitting model is defined as the set of parameters $\mathbf{a}$ that minimizes (2.22).

One method of minimizing χ^2 is to use the second-order Taylor expansion of (2.22) in terms of the parameter set $\mathbf{a}$

$$\chi^2(\mathbf{a}) \approx \gamma - \mathbf{d} \cdot \mathbf{a} + \frac{1}{2}\mathbf{a}^t \mathbf{D} \mathbf{a}, \tag{2.23}$$

where $\gamma \in \mathbb{R}$, $\mathbf{d} \in \mathbb{R}^m$, and $\mathbf{D} \in \mathbb{R}^{m \times m}$ represent the zeroth, first and second order terms of the expansion. If one knows a set of parameters $\mathbf{a}^i$ near the function minimum at $\mathbf{a}^*$, the location of the minimum can be estimated using the second-order expansion (2.23) and the local gradient of χ^2

$$\mathbf{a}^* = \mathbf{a}^i + \mathbf{D}^{-1} \cdot \left[-\nabla \chi^2(\mathbf{a}^i) \right]. \tag{2.24}$$

The matrix $\mathbf{D}$ in equations (2.23) and (2.24) is composed of all the second order derivatives of χ^2 at $\mathbf{a} = \mathbf{a}^i$ and is called the Hessian matrix. The method of finding the minimum via (2.24) is referred to as the inverse Hessian method, and is accurate when one has a set of parameters $\mathbf{a}$ in the neighborhood of the minimum. However, when $\mathbf{a}^i$ is not near the minimum, the inverse Hessian method may lead to a poor local approximation of the minimum.

If the inverse Hessian method can not give a good local estimation to the minimum, the popular steepest descent (ascent) method can provide an updated guess

of the parameters $\mathbf{a}^{i+1}$ in the direction of the minimum [143]. The direction toward the minimum value is approximated with the local gradient of χ^2 at $\mathbf{a_i}$ and a step size in this direction is chosen to update the current guess of the non-linear model parameters. Specifically,

$$\mathbf{a}^{i+1} = \mathbf{a}^i - \kappa \nabla \chi^2(\mathbf{a}^i),$$

where the constant κ is the step size in the direction of the minimum and is small enough so as not to exhaust the downhill direction.

The clever idea behind the Levenberg-Marquardt method is the seamless blending of the two minimization techniques: the inverse Hessian method and steepest descent method. The goal is to rely on the steepest descent method when the current guess for $\mathbf{a}^i$ is thought to be far from the minimum, then switch to the inverse Hessian method near the minimum. The Levenberg-Marquardt method combines both minimization techniques into a single iteration with a non-dimensional correction factor ϕ used to measure the relative strength of each method's contribution. When $\phi \ll 1$, the inverse Hessian method is dominant; and when $\phi \gg 1$, the steepest descent method dominates. At each new iteration $\mathbf{a}^{i+1}$, the merit function $\chi^2(\mathbf{a}^{i+1})$ is calculated. If the update improves $\chi^2(\mathbf{a}^{i+1}) < \chi^2(\mathbf{a^i})$, the factor ϕ is reduced to make the contribution of the inverse Hessian method more important. If the update does not improve $\chi^2(\mathbf{a}^{i+1}) \geq \chi^2(\mathbf{a}^i)$, the update is discarded and the factor ϕ is increased to make the contribution of the steepest descent method more important. The updating process continues until a suitable stopping criterion is reached. Exact details for implementing the Levenberg-Marquardt method are given in [147].

2.5.2 Model sensitivity

Previous constructions of unified models for Poiseuille and Couette flows have not discussed the sensitivity of the models at various Knudsen numbers [69, 15, 12, 11]. A more recent evaluation [112] of the unified models discusses the increased sensitivity at higher Knudsen numbers. At large Knudsen numbers, this sensitivity is shown to produce large errors when the unified models are used to predict non-equilibrium flows from a different database than used in their construction. It is crucial to understand the effect of sensitivity on producing accurate unified models and evaluating their range of applicability. In this investigation, a concrete measure of the model sensitivity is developed and used to minimize the effective error in the unified models for all degrees of rarefaction.

The unified model construction in [112] used an *ad hoc* graphical search to find the non-linear parameters that minimize the model's fit to the optimum coefficients. The observed sensitivity at high Knudsen numbers coupled with scatter found at low Knudsen numbers prompted this author to exclude the data from $Kn \leq 0.02$ in the previous non-linear model constructions. For this investigation, new unified models for Couette and Poiseuille are found using the more elegant Levenberg-Marquardt method. Furthermore, the sample data is weighted using the σ_i terms in (2.22). If the accuracy of the continuum correction is very sensitive to the choice of C_s and C_μ at a specific Kn, then σ_i is chosen to be smaller than average to increase the importance of matching the model to the specific datum. Conversely, if the choice of C_s and C_μ at a specific Kn does not have a pronounced effect on the accuracy, then σ_i is chosen to be larger than average. The basic idea of the new construction is to generate a reasonable approximation to the models sensitivity and weight the data accordingly to obtain a model with uniform accuracy for all Knudsen numbers

tested.

In general, the non-linear model can not exactly match all the data points. To determine which data points should be matched the closest, the sensitivity of each measured flow parameter to changes in the correction factors (C_s and C_μ) is found. Given an arbitrary flow measurement P, the dependence of a coefficient C from the unified model (either C_s or C_μ) can be approximated by a first-order Taylor expansion around the optimum coefficient C^* found in Table 2.2

$$P(C) \approx P(C^*) + \left.\frac{\partial P}{\partial C}\right|_{C=C^*} \cdot (C - C^*).$$

The sensitivity of the flow parameter P to a change $\Delta C = (C - C^*)$ is defined by the value of the first derivative in the Taylor series evaluated at $C = C^*$

$$\begin{aligned} \Delta P = P - P(C*) &\approx \left.\tfrac{\partial P}{\partial C}\right|_{C=C^*} \cdot \Delta C \\ |\Delta P| &\approx |G||\Delta C| \end{aligned}$$

The sensitivity $|G|$ is an estimation of the magnitude of change in the flow parameter $|\Delta P|$ when there is a change in the continuum correction $|\Delta C|$. If the sensitivity of the Navier-Stokes solution $|G(Kn)|$ is much larger than average at a specific datum, then it is more important that the model captures this point. In order to reflect this design ideal, the merit function χ^2 (2.22) to be minimized when searching for the best model is weighted by the sensitivity

$$\sigma_i = |G(Kn_i)|^{-1}.$$

For Couette flow, the unified model coefficients C_s and C_μ affect both the velocity profile and the shear stress. To quantify the effect of changing these coefficients, the first order Taylor expansions of the Navier-Stokes solutions are found for the non-dimensional slip velocity V_s and the non-dimensional shear stress T_{xy}. The slip

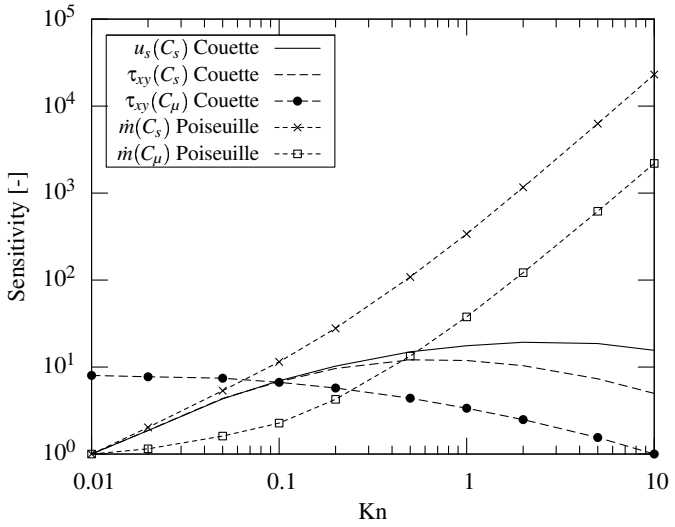

Figure 2.4: Sensitivity of the model coefficients C_s and C_μ.

velocity is obtained by normalizing the solution of (2.16) at the wall $\eta = 0$ by half the difference in the wall speeds $\frac{1}{2}(U_1 - U_0)$

$$V_s(C_s) = \frac{2U_0}{U_1 - U0} + \frac{2C_s Kn}{1 + 2C_s Kn}. \tag{2.25}$$

Note the slip velocity of the corrected Navier-Stokes solution only depends on C_s. The sensitivity of the Couette slip velocity to the slip coefficient C_s is found from the first-order Taylor expansion of (2.25) around the point $C_s = C_s^*$

$$|G[u_s(C_s)]| = \frac{2Kn}{(1 + 2C_s^* Kn)^2}. \tag{2.26}$$

The optimum values of C_s^* are taken from Table 2.2. In Figure 2.4, the sensitivity of the slip velocity to changes in the slip coefficient is found to increase with Knudsen number by an order of magnitude over the range $0.01 \leq Kn \leq 10$. All the sensitivities presented in Figure 2.4 are normalized by the minimum value found in the range $0.01 \leq Kn \leq 10$. Any changes this normalization makes to the importance weighting σ_i can be factored out of the merit function χ^2, and thus have no influence on the Levenberg-Marquardt minimization technique.

The Navier-Stokes solution to the Couette flow shear stress (2.18) is normalized by the continuum shear stress $\tau_{xy} = \mu_0(U_2 - U_1)/h$

$$T_{xy}(C_s, C_\mu) = \frac{C_\mu}{1 + 2C_s Kn}.$$

The sensitivity of the shear stress to changes in the unified model coefficients is found by the first-order Taylor expansion in each coefficient around its optimal value

$$|G[\tau_{xy}(C_s)]| = \frac{2C_\mu^* Kn}{(1 + 2C_s^* Kn)^2}, \tag{2.27}$$

and

$$|G[\tau_{xy}(C_\mu)]| = \frac{1}{1 + 2C_s^* Kn}. \tag{2.28}$$

Similar to the slip velocity, the sensitivity of the shear stress to the slip coefficient C_s varies by an order of magnitude, in Figure 2.4. The sensitivity $|G[\tau_{xy}(C_s)]|$ increases with Knudsen number until a maximum factor of 12 is reached at $Kn = 0.5$, then decreases to a factor of 5 at $Kn = 10$. Since there are two sensitivity estimates for the slip coefficient (2.26) and (2.27), only the sensitivity associated with the slip velocity $|G[u_s(C_s)]|$ (2.26) is used for the importance weighting σ_i of the merit function χ^2. The sensitivity of the shear stress to the viscosity correction C_μ decreases by nearly an order of magnitude with increasing Knudsen number. As the only measure of sensitivity for the viscosity correction affect on Couette flow, $|G[\tau_{xy}(C_\mu)]|$ in (2.28) is used for the importance weighting σ_i of the merit function χ^2.

For Poiseuille flow, the unified model coefficients C_s and C_μ both affect the velocity profile. The sensitivity is measured for the non-dimensional M, defined as the velocity in (2.17) averaged over the channel and normalized by the quantity $\rho f h^2/\mu_0$

$$M(C_s, C_\mu) = \frac{1}{2C_\mu}\left(C_s Kn + \frac{1}{6}\right). \tag{2.29}$$

Note the normalization of the average velocity is such that any error in (2.29) is equivalent to the normalized error in mass flux and number flux. The sensitivity of the average velocity to the unified coefficients is found again from the first-order Taylor expansion of (2.29) around the optimum coefficients C_s^* and C_μ^*

$$|G[\dot{m}(C_s)]| = \frac{Kn}{C_\mu^*},$$

and

$$|G[\dot{m}(C_\mu)]| = \frac{1}{2C_\mu^{*2}}\left(C_s^* Kn + \frac{1}{6}\right).$$

In Figure 2.4, the sensitivity of the average velocity increases monotonically with Knudsen number for both coefficients. The sensitivity to the slip coefficient C_s is over 20,000 times larger at $Kn = 10$ than at $Kn = 0.01$. Similarly, the sensitivity to the viscosity correction C_μ is 2,200 times larger at $Kn = 10$ than at $Kn = 0.01$. The contrast in the variation of sensitivity between the two flow types demonstrates that unified Couette flow models tend to yield better results for more non-equilibrium flows because the errors are more "forgiving" at higher Knudsen numbers than Poiseuille flow errors.

2.5.3 Candidate models

A total of four non-linear models is tested to find which functional form best captures the optimum slip coefficient C_s for argon Couette flow. All models rely on a monotonically decreasing function with an asymptotic limit as $Kn \to \infty$. The first model is referred to as an arctangent law and adopts the same form as the four parameter BPB model for the slip coefficient (2.13)

$$C_s(Kn) = a_1 - a_2 \tan^{-1}(a_3 Kn^{a_4}). \tag{2.30}$$

The second model, called the power law, is also defined by four parameters but the arctangent function is replaced with an offset power law relationship

$$C_s(Kn) = a_1 + \frac{a_2}{(Kn + a_3)^{a_4}}. \tag{2.31}$$

The two remaining models only use three parameters in an effort to see if a slightly simpler non-linear form could be found by eliminating the power scaling of a_4 in (2.30) and (2.31). The Levenberg-Marquardt method can detect when the power scaling is unnecessary in either of the previous models $a_4 \approx 1$, thus the three parameter models use different monotonic functions. They include an exponential decay law

$$C_s(Kn) = a_1 + a_2 e^{-a_3 Kn},$$

and a hyperbolic tangent law

$$C_s(Kn) = a_1 - a_2 \tanh(a_3 Kn).$$

The Levenberg-Marquardt method as presented in [147] is used to find the best fitting non-linear model to the optimum slip coefficient for argon Couette flow. The model errors in the merit function (2.22) are weighted using the inverse of the gain for the slip coefficient found in (2.26)

$$\sigma_i = |G[u_s(C_s)]|^{-1}. \tag{2.32}$$

Models with coefficients used as exponents tend to produce very unstable iterations with the Levenberg-Marquardt method, even with a drastic reduction (10^{-3}) in the update step size. As a quick stability fix, the exponent parameter is held constant through the Levenberg-Marquardt iteration. Once the method converges to the best fitting model for a given exponent, a new exponent is calculated using the bisection method in the neighborhood of the single global minimum of the merit function χ^2.

Name	Model Form	a_1	a_2	a_3	a_4	$\chi^2(\mathbf{a})$
Arctan Law	$a_1 - a_2 \tan^{-1}(a_3 Kn^{a_4})$	1.546	0.869	1.357	0.545	1.352
Power Law	$a_1 + \frac{a_2}{(Kn+a_3)^{a_4}}$	0.180	0.631	0.286	0.532	1.444
Exp Law	$a_1 + a_2 e^{-a_3 Kn}$	0.408	0.821	0.874	-	3.456
Tanh Law	$a_1 - a_2 \tanh(a_3 Kn)$	1.178	0.760	0.615	-	5.002

Table 2.3: Candidate non-linear model laws for the Couette flow slip coefficient for argon gas.

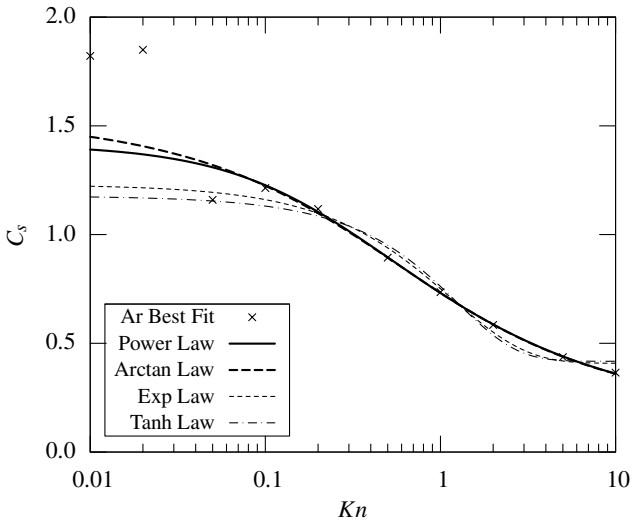

Figure 2.5: Comparison of the four non-linear models ability to capture the optimum slip coefficient for the Couette flow of argon gas.

With the new exponent parameter fixed, the Levenberg-Marquardt method is used again to calculate the best-fitting model. The process is repeated until the exponent parameter corresponding to the global minimum of χ^2 is found.

The best fitting coefficients found with the fixed-exponent Levenberg-Marquardt method and the corresponding merit function χ^2 are presented in Table 2.3. The four parameter models fit the optimum C_s two to three times better than the three parameter models. A visual comparison of all four models to the actual optimum coefficients is given in Figure 2.5. It is clear that the three parameter models are too "stiff" to accurately capture the shape of the Knudsen number dependence of

C_s. The power scaling offered by the coefficient a_4 in the offset power law (2.31) and arctangent law (2.30) is needed to appropriately stretch the non-linear model in the Kn range to fit the curve at the larger Knudsen numbers. The error sensitivity of the slip coefficient is ten times greater for $Kn \geq 0.5$ than for $Kn = 0.01$. Using the sample weighting proposed (2.32) yields nearly exact matches of the optimum coefficient for both the power law and arctangent law for $Kn \geq 0.5$. The only noticeable difference in the four parameter models occurs for $Kn < 0.05$ when the sensitivity of the slip coefficient is comparatively small.

2.5.4 Model Selection

Given the success of the non-linear models shown in Figure 2.5, only the arctangent law (2.30) and the offset power law (2.31) are considered candidates for the remaining model selection. The best-fitting model is found for the optimum slip coefficient C_s^* and viscosity correction C_μ^* for both Couette and Poiseuille argon gas flows from Table 2.2. The fixed-exponent Levenberg-Marquardt method described in the previous section is used to find the best-fitting parameters to the proposed non-linear models. The merit function (2.22) is weighted using the appropriate sensitivity calculations from Section 2.5.2. It is important to recall that the slip coefficient C_s for Couette flow has two sensitivities; one corresponding to its effect on the slip velocity and another for its effect on the average shear stress. While the two sensitivities are similar in Figure 2.4, only the sensitivity derived from the slip velocity is used because the slip velocity is solely influenced by the slip coefficient.

The best-fitting non-linear models are found for matching the optimum slip coefficient C_s^* and viscosity correction C_μ^* in Figure 2.6. In a previous study [112], the optimum coefficients for both argon and nitrogen gas flows are combined to generate

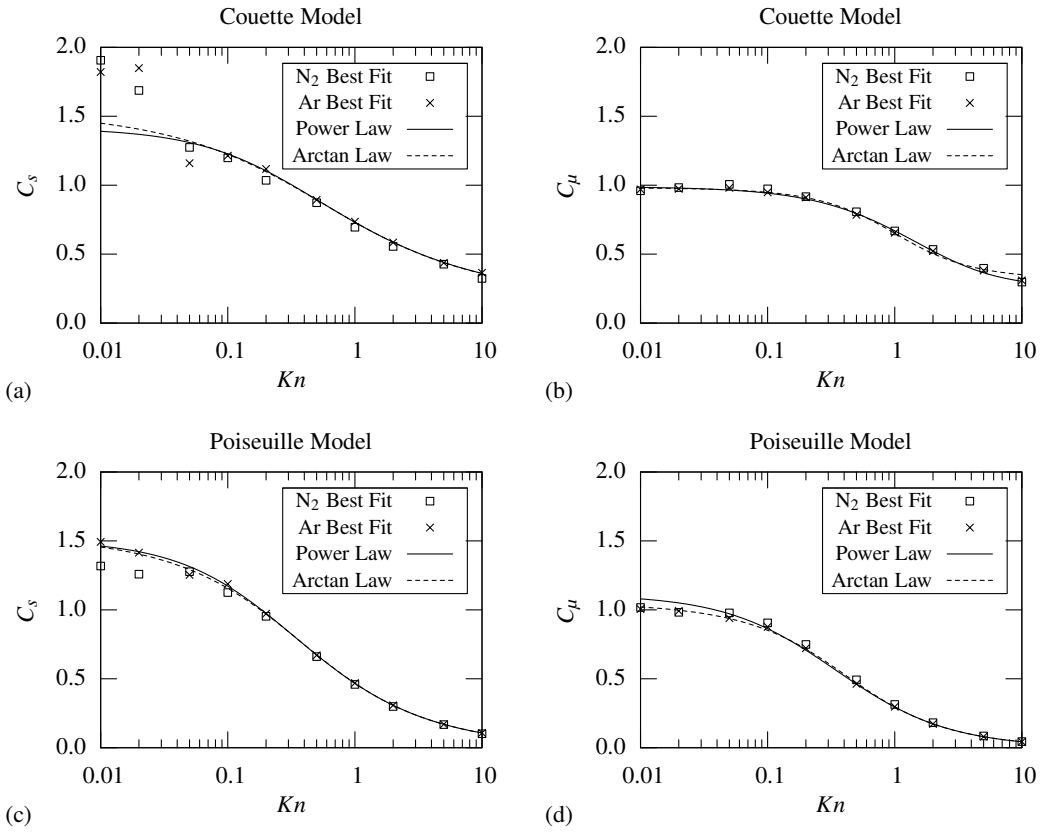

Figure 2.6: Non-linear model construction for the optimum continuum corrections for argon gas flows: (a) Couette flow slip coefficient C_s; (b) Couette flow viscosity correction C_μ; (c) Poiseuille flow slip coefficient C_s; and (d) Poiseuille flow viscosity correction C_μ.

a single unified model for both gases. This is motivated in part because the optimum coefficients for the two gases are within 5% for almost every case when $Kn \geq 0.1$. However, even a 0.1% difference between the coefficients could result in an error of over 10% for Poiseuille flows at high Knudsen numbers (*i.e.* $Kn \geq 1$). From the results on the model sensitivity in Figure 2.4, this is to be expected. Therefore, the new unified models constructed in this investigation will only use the optimum coefficients for the argon cases.

For the Couette flow cases, the power law is found to give the best overall fit to both sets of optimum coefficients. In fact, the optimum exponents in the power law are sufficiently close to aesthetically appealing values $\{1/2, 2\}$ that the cleaner form is adopted while accepting a minimal (0.1%) increase in the merit function. The new unified model proposed in this investigation for argon Couette flow is

$$\begin{aligned} C_s(Kn) &= 0.161 + \frac{0.641}{\sqrt{Kn + 0.262}} \\ C_\mu(Kn) &= 0.266 + \frac{6.288}{(Kn + 2.949)^2}. \end{aligned} \tag{2.33}$$

For the Poiseuille flow case, the arctangent law is found to give the superior fit for both optimum coefficients. The new unified model proposed in this investigation for argon Poiseuille flow is

$$\begin{aligned} C_s(Kn) &= 1.543 - 0.983 \tan^{-1}(1.935 Kn^{0.669}) \\ C_\mu(Kn) &= 1.051 - 0.671 \tan^{-1}(2.091 Kn^{0.835}). \end{aligned} \tag{2.34}$$

2.5.5 Model Error

The new unified models for Couette (2.33) and Poiseuille (2.34) flows are tested against the non-equilibrium cases in the DSMC database. In addition to the new models, the BPB model (2.13) and (2.14) for Couette flow and the KB model (2.10)

and (2.11) are also tested against the database. It is important to note that the accuracy shown by the models in these next examples is not a measure of their *predictive* power. Instead, it is a measure of how close a corrected Navier-Stokes solution can match a non-equilibrium solution and how close can the Knudsen number dependence of those corrections can be approximated. The non-equilibrium solutions from the database are used to construct accurate corrections to the Navier-Stokes solution through the unified models. Thus, the unified models are not really predicting the non-equilibrium results rather they are only reflecting the accuracy of the data fit.

The error measures used for comparison are the L_2 error (2.21) of the Couette and Poiseuille velocity profiles and the average shear stress τ_{xy} throughout the Couette channel. The error in slip velocity is not used because it is found to closely track with the L_2 error in the velocity profile [112]. The error in the Poiseuille flow mass flux is not used because it is bounded from above by the L_2 error in the velocity profile and it too closely tracks the L_2 error. The error in wall shear stress is not used for Poiseuille flow because it only depends on the density, body force and channel height (2.19). Since both the DSMC method and the corrected Navier-Stokes equation conserve momentum in detail, the wall shear stress is independent of the method used for its calculation. Thus, the wall shear stress will be identical and is not useful in evaluating these models.

Both the new unified Couette model (2.33) and the BPB model recover the non-equilibrium Couette velocity profiles to within an L_2 error norm of 2.5% for $0.01 \leq Kn \leq 10$ as illustrated in Figure 2.7. Neither case has an error more than 1% larger than the best possible error. The best possible error is the minimum L_2 error found by the linear least squares method when finding the optimum coefficients. Furthermore, the new model performance appears to be independent of operating

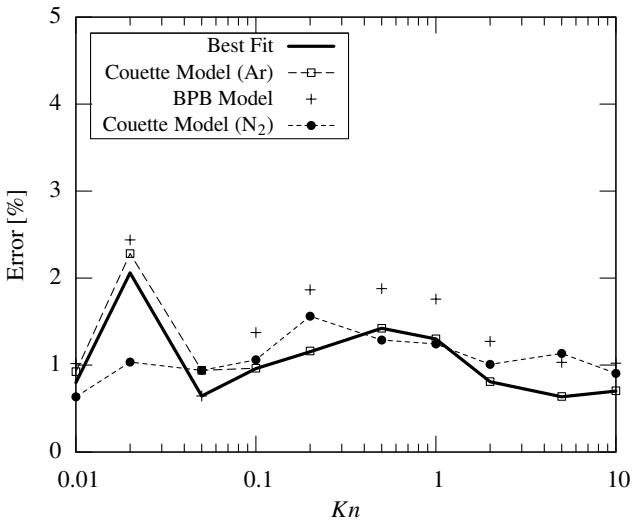

Figure 2.7: L_2 error in the velocity profiles of the continuum corrections for Couette flow.

fluid when predicting the velocity profile. The new unified model constructed solely from non-equilibrium argon cases produces a similar error for the non-equilibrium nitrogen cases for $0.01 \leq Kn \leq 10$.

The new unified model (2.33) recovers the average Couette channel shear stress to within 5% for $0.01 \leq Kn \leq 10$ as illustrated in Figure 2.8. Compared to the unified model proposed in [112], the new model is almost 2.5 times more accurate. For argon Couette flows, using the sensitivity to weight the non-linear model error has effectively distributed the error across all Knudsen numbers in the database. The non-equilibrium nitrogen cases using the new unified model predominantly have a larger error than the argon cases with a maximum error of over 10%. The difference in shear stress error between the two gas types indicates that a unified model is most accurate for the cases from which it is derived. The BPB model shows a considerable error in the average shear stress at low Knudsen numbers with a maximum error of 18%. Does this mean that the new unified model is superior to the other models

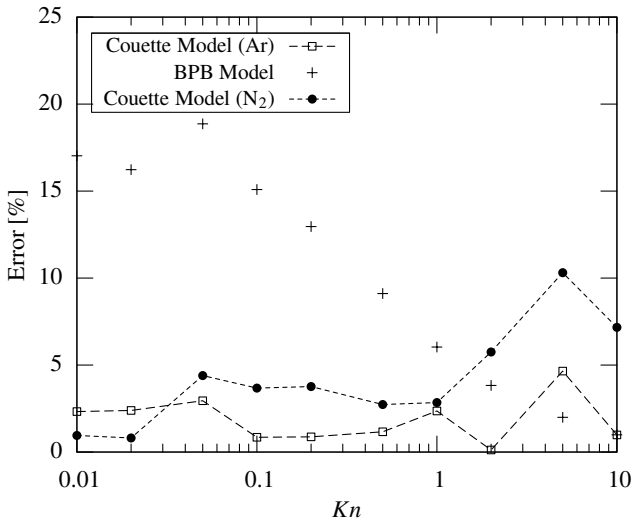

Figure 2.8: Relative error in the average shear stress τ_{xy} of the continuum corrections for Couette flow.

in the literature? No, it is only superior on its own database and simply further demonstrates the sensitivity of the unified models to their construction schemes and assumptions. The BPB model is constructed from a database of linearized Boltzmann solutions and reports a shear stress error of only 0.3% when compared to its own database [12].

The new unified model recovers the non-equilibrium Poiseuille velocity profile to within an L_2 error of 2% for $0.01 \leq Kn \leq 10$ which is within 1% of the best possible L_2 error norm. The nitrogen cases using the new unified model show a maximum error of 8%, and for all cases when $Kn \geq 0.05$, the error is at least twice as great as the argon cases. The model sensitivity to the operating conditions of the database at high Knudsen numbers is illustrated by the difference between gas species. It indicates that any unified Poiseuille model should only be applied to the working fluid represented in the database used to design the model for large Knudsen numbers. The KB model has a maximum L_2 error of 20% in its Poiseuille velocity

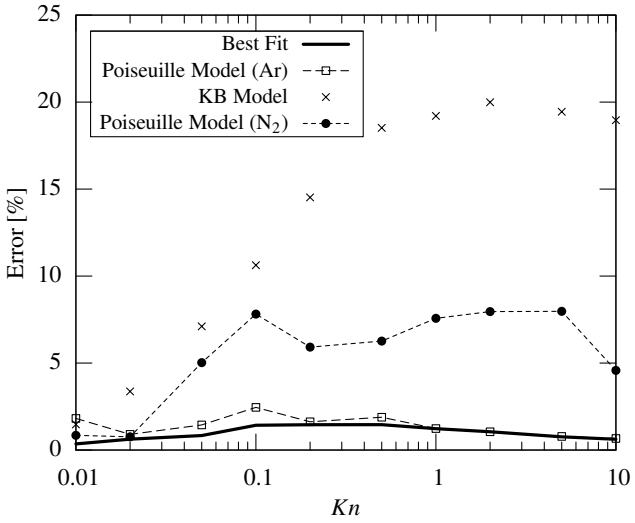

Figure 2.9: L_2 error in the velocity profiles of the continuum corrections for Poiseuille flow.

profile. Again, this drastic departure in performance at large Knudsen numbers is due to the differences in the non-equilibrium databases used to construct the models presented here.

2.6 Empirical Model Prediction Performance

The investigation demonstrates that it is possible to construct models for the slip and viscosity coefficients in such a manner that the corrected Navier-Stokes solutions accurately capture the non-equilibrium solutions in the DSMC database. However, this is only a measure of the quality of the data fit the unified models are able to achieve, and is not a testament to their use as a predictive design tool. In order to understand the ability of a model constructed from a database of non-equilibrium solutions to predict other flows outside the database, the following five cases are tested: interpolation and extrapolation of the DSMC cases, combined Couette and Poiseuille flow, wall surfaces with partial momentum accommodation, helium gas flows, and channel flows with uniform suction and injection. In each case, only the Couette and

Poiseuille models developed in this investigation are compared to the DSMC results. In Figures 2.1 and 2.2, the velocity profile of the Navier-Stokes solution using the new unified models (solid line) is presented with the non-equilibrium solution from the DSMC database for argon Couette and Poiseuille flows at $Kn = 0.01$, 0.1, 1, and 10. As a reference for the quality of the predictions, the results in Figure 2.1 demonstrate the ability of our Couette model to capture the velocity profile of the DSMC Couette solution to within an L_2 error norm of 2%, and a shear stress error of 4%. Similarly for Poiseuille flow in Figure 2.2, the new unified Poiseuille model is able to capture the velocity profile of the DSMC solution to within an L_2 error norm of 2%.

2.6.1 Interpolation and Extrapolation

Couette and Poiseuille argon gas flows are predicted using the models at $Kn = 0.7$ and $Kn = 20$, which is an interpolation and extrapolation of the cases used in the databases. For the interpolation case $Kn = 0.7$, the models predict all measurable error quantities within the baseline database accuracy. This indicates that an empirical model can serve as a tool to evaluate different operating densities of the same geometry over a wide range of Knudsen numbers, if there are enough non-equilibrium solutions available to construct the model. The number of non-equilibrium cases required in the database depends on the fidelity hoped to be achieved by the model and the complexity of the Knudsen number dependence of the system. For the extrapolation case $Kn = 20$, the velocity profiles of both flow types are within the baseline 2% accuracy. However, the average shear stress predicted by the new Couette model is in error by 30%. This is a consequence of selecting a purely empirical model form in (2.33). The BPB model (2.14), which is designed specifically to recover the

asymptotic value in the free molecular limit, predicts the shear stress to 0.01%.

2.6.2 Combination of Couette and Poiseuille Flow

The simplified Navier-Stokes equations for the Couette (2.9) and Poiseuille (2.8) flows considered in this investigation each reduce to a single linear differential equation for the velocity profile. Therefore, in the continuum limit, a flow that is a combination of Couette and Poiseuille flows can be solved as the superposition of a Couette solution and a Poiseuille solution. However, the Boltzmann equation, which is valid for flows ranging from continuum to free molecular, has a nonlinear collision term which prevents the linear superposition of the two flows in the transition regime. If the velocity distribution function within the flow is still close to equilibrium, then the error due to the non-linearity is small. In order to evaluate the effect of the non-linearity, a combination Couette and Poiseuille argon flow at $Kn = 1$ is tested. The Navier-Stokes result is obtained by decomposing the flow into a Couette and Poiseuille contribution, solving each separately with the unified Couette and Poiseuille models developed in this investigation, and then adding the two solutions together under the principle of superposition. In Figure 2.10, the DSMC solution (circles) and the Navier-Stokes solution (solid line) using our models are presented for the combined Couette and Poiseuille velocity profile at $Kn = 1$. The lower wall is fixed while the upper wall moves at 20 m/sec in the direction of the driving force. The driving force combined with the moving wall boundary yields a maximum velocity of about 30 m/sec in the DSMC solution. The Navier-Stokes solution predicts the entire velocity profile resulting to within an L_2 error of 1%, which is the same as the reference Poiseuille case at $Kn = 1$. Moreover, the wall shear stress error is only 2% and is less than the reference Couette flow case. It is important to note that the

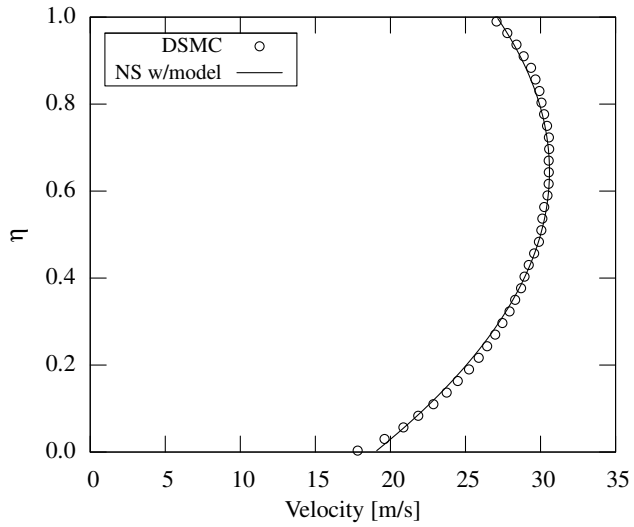

Figure 2.10: Combined Couette and Poiseuille flow for Argon gas at $Kn = 1$.

Poiseuille solution contributes 60% of the wall shear stress, and that the Poiseuille wall shear stress is identically solved by both the Navier-Stokes and DSMC method. In Figure 2.10, the largest error in the Navier-Stokes prediction of the velocity profile occurs near the wall where the slip velocity is 5% larger than the DSMC result. This slip velocity error is slightly larger than the error magnitude found in the reference Poiseuille case at $Kn = 1$. Overall, the effect of any non-equilibrium non-linearity appears small, and the decomposition of two flow types is appropriate in this case.

2.6.3 Tangential Momentum Accommodation Coefficient

In order to determine the effect the TMAC has on the models' performance, argon gas Couette and Poiseuille flows are simulated for a TMAC equal to 0.8 and 0.5 at $Kn = 1$. For the Couette flows, the TMAC has no effect on the accuracy of the velocity profile, with the Couette model predicting the profiles to the same accuracy as the baseline. However, the shear stress error triples to 6% for a TMAC = 0.8 and is 5 times larger for a TMAC = 0.5. In Figure 2.11, the DSMC solution (circles)

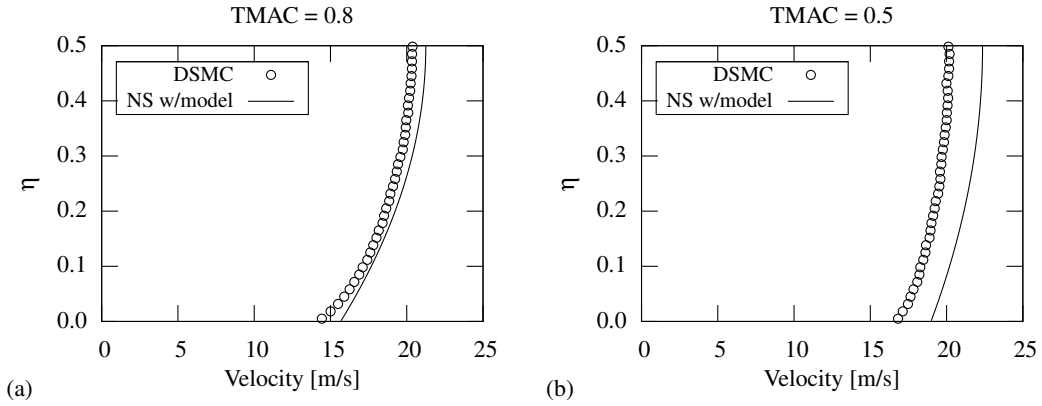

Figure 2.11: Poiseuille flow for Argon gas with different TMACs at $Kn = 1$: (a) TMAC = 0.8; and (b) TMAC = 0.5.

and the Navier-Stokes solution (solid line) using the new unified Poiseuille model are presented for the Poiseuille velocity profiles at $Kn = 1$, with the TMAC equal to 0.8 and 0.5. For $\sigma_v = 0.8$, the L_2 error between the velocity profiles predicted by DSMC and the Navier-Stokes equation is twice as large as the reference Poiseuille case at $Kn = 1$. However, as shown in Figure 2.11, the Navier-Stokes prediction worsens when the TMAC equals 0.5. In this case, the Navier-Stokes solution with the new Poiseuille model over-predicts the velocity across the entire channel with an error five times that found in the reference Poiseuille case at $Kn = 1$. The deviation from equilibrium is intricately coupled with the range of direct influence of the wall on the gas molecules of the flow. As the Knudsen number increases, so does the probability of finding a molecule whose last collision was with the wall. It is reasonable to assume that the TMAC will be a sensitive factor at higher Knudsen numbers. Therefore, care should be exercised when using an empirical model to predict a flow with a TMAC value different from that used in the database to construct the model.

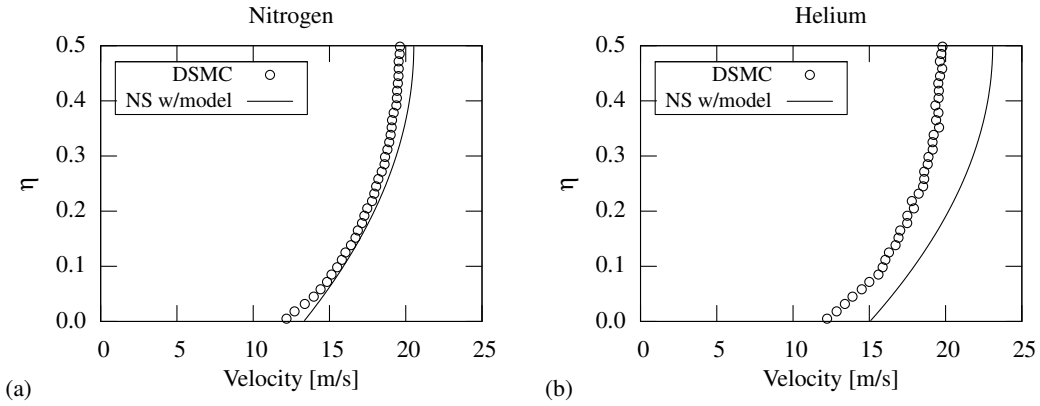

Figure 2.12: Poiseuille flow for different gases at $Kn = 1$ (a) Nitrogen and (b) Helium.

2.6.4 Helium Gas Flows

The molecular weight of helium is one tenth that of argon, which means the most probable molecular speed of helium is over 3 times faster than argon. By comparison the difference between the most probable molecular speeds of argon and nitrogen is only 16%. While the optimum coefficients found in the database for each gas (Table 2.2) are similar, the errors at high Knudsen numbers are larger when the argon model is used for nitrogen gas flows. The large difference in molecular speeds between the helium cases and the models' database could affect the accuracy of the models' prediction. In order to evaluate this large change in molecular speeds, helium Couette and Poiseuille flows are simulated at $Kn = 1$. The helium gas is found to have no effect on the ability of the new unified Couette model to predict Couette velocity profile, the L_2 error is within the model accuracy of 2% for the baseline case. In Figure 2.12, the DSMC solution (circles) and the Navier-Stokes solution (solid line) using our Poiseuille model are presented for the Poiseuille velocity profile of helium gas at $Kn = 1$. The increase in the random or thermal speed due to the lighter helium gas introduces more statistical scatter in the DSMC solution than the

argon and nitrogen cases. The increased scatter in the velocity profile is illustrated in Figure 2.12, but the overall scatter is still less than 3% across the channel. For the helium Poiseuille flow, the Navier-Stokes solution over-predicts the DSMC velocity profile throughout the channel resulting in a 12% higher mass flux than the DSMC solution. Furthermore, the error in the L_2 norm of the velocity profile is 15%, which is over 7 times larger than the reference Poiseuille case at $Kn = 1$. The absence of error in the Navier-Stokes prediction for Couette flow is due to the viscosity independence of the velocity profile and further differentiates the model performance between flow types.

2.6.5 Body force driven flow with uniform rates of suction and injection

In order to test a one dimensional flow with a non-zero convective acceleration, the boundary conditions for a body force driven flow are changed to include a uniform fluid injection at the lower wall and a uniform suction at the upper wall. In Figure 2.13, the DSMC solution (circles) and the Navier-Stokes solution (solid line) using the new unified Poiseuille model are presented for the velocity profile of a body force driven flow with uniform rates of suction and injection for argon gas at $Kn = 1$. The body force is chosen to drive the flow at a maximum velocity of 20 m/sec, while the injection and suction rates maintain a constant 20 m/sec cross flow. The presence of cross flow in the solution skews the normally symmetric Poiseuille velocity profile in the direction of the cross flow. This asymmetry in the DSMC solution creates a 10% difference between the slip velocity at the upper and lower wall boundaries, and shifts the location of the maximum velocity from the center by 8% of the channel width. The asymmetry in the Navier-Stokes solution due to the cross flow is not as pronounced as in the DSMC solution. As illustrated in Figure 2.13, the difference

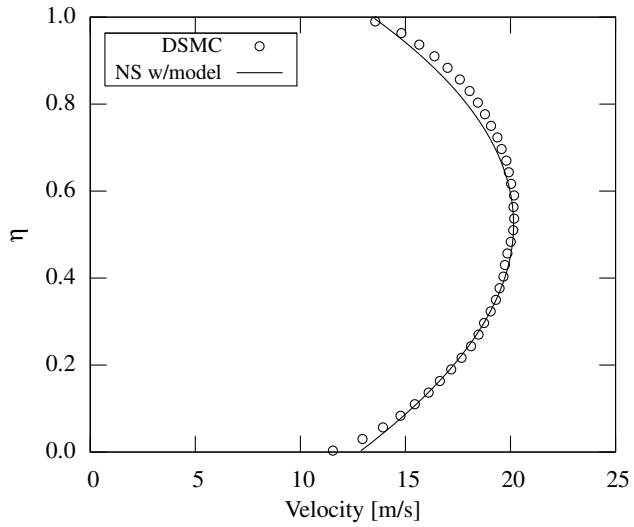

Figure 2.13: Force driven duct flow with uniform suction and injection at the walls for Argon gas at $Kn = 1$.

between the upper and lower slip velocities is less than 7% in the Navier-Stokes solution, while the location of the maximum velocity has shifted only 2% of the channel width. Furthermore, the difference in the velocity gradient at the walls between the DSMC and Navier-Stokes solutions creates an error in the wall shear stress of 8%. The Navier-Stokes solution using the new unified Poiseuille model is able to predict the mass flux to within 1% of the DSMC solution, in spite of missing the key changes in the shape of the velocity profile due to the cross flow. The Poiseuille model does not appear to follow the change of physics for the non-zero convective acceleration. Thus, it is not recommended to use the Poiseuille model developed in this investigation for complex flows when the Knudsen number is larger than 0.1.

2.7 Summary

There were two goals in this investigation. The first was to construct new unified Couette and Poiseuille models based on empirical corrections in order for the

Navier-Stokes solution to match a wide range of known non-equilibrium flows. The second was to evaluate their effectiveness as a predictive design tool. A database of non-equilibrium solutions was first simulated with DSMC for Couette and Poiseuille flows of argon and nitrogen, for Knudsen numbers ranging from 0.01 to 10. Then the optimum slip and viscosity model coefficients for the Navier-Stokes solution were found for each flow condition so that the velocity profile and shear stress matched the DSMC data in a linear least squares sense. Next, models were constructed for each flow type in order to capture the Knudsen number dependence of the slip and viscosity model coefficients. The new unified Couette and Poiseuille models developed in this investigation demonstrated their ability to capture all the non-equilibrium results in the DSMC database for Couette flows with an L_2 error norm in the velocity profile of 2% and a shear stress error of 5%. Similarly for Poiseuille flows, the Poiseuille model captured the results in the DSMC database with all velocity and mass flux errors within 2%. The performance of the Couette and Poiseuille models developed here is similar to other unified models proposed by Beskok and Karniadakis, and Bahukudumbi, Park, and Beskok. All models, even those purposely used on flows that were not their intended design, were accurate for near equilibrium conditions at Knudsen numbers less than or equal to 0.1. Above this Knudsen number, the correction to the viscosity model coefficient indicates that the error in the shear stress closure is at least 10% for Couette flows and 20% for Poiseuille flows. The models' performance capturing the DSMC database was very sensitive in the transition and free molecular regimes. Generally, as the Knudsen number increases, so does the error using any model that was not explicitly constructed from the database used in the comparison. The new unified Couette and Poiseuille models developed in this investigation were able to predict flows that are an interpolation of the DSMC

database to a similar accuracy as the database cases themselves. However, a lack of asymptotic information in the Couette shear stress correction led to a 30% error for the extrapolation case $Kn = 20$. In addition, a combination of both models was able to predict a combined Couette and Poiseuille flow in the transition regime. The Couette model was successful in predicting the velocity of all the cases because the Navier-Stokes solution is independent of any errors in the shear stress closure due to non-equilibrium. However, the Poiseuille model was not as successful in predicting flows with partial wall accommodation, helium gas, and non-zero convective acceleration terms. The models developed in this investigation are empirical corrections to a continuum solution that has little physical accuracy in the transition and free molecular regimes, and the errors found when pushing the models outside the database in these regimes are expected.

CHAPTER III

QUASI-MONTE CARLO CONVERGENCE

Low speed, non-equilibrium gas flows represent one of the most challenging fluid simulation problems. As noted in Section 1.2.1, traditional CFD techniques based on near continuum approximations fail to accurately simulate non-equilibrium flows because the no-slip boundary condition and transport closures are no longer valid. This loss of validity is attributed to the deviation of the flow from local thermodynamic equilibrium. There is an insufficient number collisions occurring in the near wall region for the flow to relax to the wall conditions. Similarly, there is an insufficient number collisions occurring throughout the flow to represent the continuum transport of mass, momentum, and energy. The DSMC method of Bird [16] is a particle method that is able to obtain physically accurate non-equilibrium solutions by actually simulating the probabilistic behavior of the gas based on kinetic theory. While accurate, the DSMC method suffers from long simulation times compared to the Navier-Stokes solutions (see Section 1.2.2). Moreover, when the average velocity of the flow in interest becomes sufficiently small relative to the random fluctuations associated with the thermal energy of the gas, the DSMC method becomes computational intractable in practice.

The goal of this investigation is to try and obtain an accurate and efficient simu-

lation of low speed, non-equilibrium gas flows for micro-scale applications. The first approach of this investigation (see Chapter II) provides empirical corrections to the no-slip boundary condition, and the continuum shear stress closure for the Navier-Stokes equations to be applied to non-equilibrium flows. Because non-equilibrium solutions are needed *a priori* for such corrections, the empirical approach has unsatisfyingly limited predictive capabilities. The second approach of this investigation, which is covered in the remaining chapters, is to develop a quasi-Monte Carlo (QMC) particle technique. QMC is an approximate integration technique that uses the same framework as the Monte Carlo method to obtain an estimate by averaging samples of the integrand. The appeal of the QMC technique is that, in theory, the method enjoys a near linear error convergence rate with the number of samples. In contrast, the Monte Carlo method converges in $\mathcal{O}\left(N^{-1/2}\right)$ time with N samples. It is this slow error convergence rate that impedes the application of the DSMC method to low speed, non-equilibrium flows. Thus, a QMC particle method has the potential to retain the physical accuracy of the DSMC simulation, while computing the desired solution in significantly less time than DSMC.

The remainder of the chapter is devoted to the review of the basic theory of quasi-Monte Carlo integration as a means to improve the error convergence rate of DSMC. In Section 3.1, the theory concerning the error of the general Monte Carlo and QMC integration methods is presented. The anticipated performance gains associated with the QMC theory are attributed to the result of Koksma and Hlawka. Their result states that the integral approximation error of any sampling method, Monte Carlo or QMC, is bounded by two quantities, the discrepancy of the sample points, and the variation of the integrand. While the concepts of discrepancy and variation are common in real analysis, they typically are not discussed in regards

to most computational fluid techniques. For this reason, an engineering description with examples is provided in Section 3.2 for the discrepancy, and Section 3.3 for the variation. The QMC method improves the error convergence of its integral approximation by sampling the integrand with a point set with a lower discrepancy than random. In Section 3.4, two approaches for producing these low-discrepancy point sets are reviewed. Specifically, the advantages of using a low-discrepancy sequence instead of an optimal integration lattice are highlighted for applications of the QMC method in this investigation.

3.1 Integration Error

In order to understand when the computational costs associated with the DSMC method become intractable for low speed non-equilibrium flows, the general theory behind Monte Carlo integration is presented in this section. Quasi-Monte Carlo (QMC) integration attempts to improve the error convergence rate of Monte Carlo by replacing the random sample points with a "better" distribution. The theory behind QMC integration is also presented in this section; specifically, as it relates to the ability of QMC to achieve a superior error convergence rate to Monte Carlo in an asymptotic sense. Unfortunately, the integration error theory of Monte Carlo and QMC is based on the asymptotic accuracy as the number of sample points used in the approximations tends toward infinity. Moreover, the bounds associated with the integration error are not necessarily very tight. Thus, for integral approximations in practice (*i.e.* finite time and memory), one must directly compare the two methods for each specific problem to ascertain which is faster.

3.1.1 Monte Carlo Integration Error

Monte Carlo integration approximates the integral of a function by averaging randomly selected function samples taken within the integration domain. These function samples are all of equal weight. Hence, Monte Carlo integration yields a rather simple numerical approximation [47]

$$\int_{[0,1]^s} f(\mathbf{u})d\mathbf{u} \approx \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{x}_i), \text{ with } \mathbf{x}_i \in \mathscr{U}(0,1)^s, \tag{3.1}$$

where $\mathbf{x}_i \in \mathscr{U}(0,1)^s$ denotes that the vectors $\mathbf{x}_i$ are sampled from a uniform distribution taken over the unit hypercube $[0,1]^s$.

Before proceeding, a brief discussion on the integration domain and the distribution of the random variates is necessary. In the Monte Carlo approximation (3.1), the integration domain is assumed to be unit hypercube $[0,1]^s$; however, this need not always be the case. In general, Monte Carlo integration can be used on any integration domain $\mathcal{R}$, provided there exists a means to generate uniform samples on $\mathcal{R}$,

$$\int_{\mathcal{R}} f(\mathbf{u})d\mathbf{u} \approx \frac{\lambda_s(\mathcal{R})}{N}\sum_{i=1}^{N} f(\mathbf{x}_i), \text{ with } \mathbf{x}_i \in \mathcal{R},$$

where the sample average is corrected by a factor equal to the volume of the domain $\lambda_s(\mathcal{R})$. There is, however, a practical reason for limiting Monte Carlo integration to the unit hypercube. Specifically, there is no known direct means[1] to produce a random variate except from a uniform distribution on a finite interval [78]. The pseudo random number generators (PRNGs) that serve as the backbone of modern Monte Carlo integration only produce uniformly distributed variates in a finite interval $[a,b]$. Thus, any implementation of Monte Carlo integration using a PRNG must ultimately represent an integral over a finite domain $[a_1,b_1]\times[a_2,b_2]\times\cdots\times[a_s,b_s]$. It

[1]The term "direct means" used here refers to deterministic algorithms implemented on a computer and excludes the use of precomputed tables of variates, such as [32].

is then a straightforward process to apply a linear transformation to each dimension of the integration domain to map the finite intervals to the unit hypercube.

Random variates with non-uniform distributions must be obtained by a transformation of a uniform distribution (*e.g.* the inverse cumulative, or transform, method, the acceptance-rejection method, and the rectangle-wedge-tail method [78]). While a general Monte Carlo method can be analyzed mathematically for any arbitrary probability space, any computer implementation of the method must be reducible to a sampling problem using uniformly distributed variates. Therefore, the Monte Carlo approximation using uniformly distributed variates over the unit hypercube represents the vast majority of practical applications for the method. For this reason, the scope of Monte Carlo integration covered is limited to the approximation in (3.1) in this investigation. As a matter of convenience, let $\overline{I}^s$ denote the unit hypercube consisting of closed unit intervals $[0,1]^s$, and I^s denote the unit hypercube consisting of half-open unit intervals $[0,1)^s$. While the distinction between the two sets $\overline{I}$ and I^s is important to mathematicians, it has little impact on the actual implementation of the Monte Carlo method considered here.

The error in the Monte Carlo approximation (3.1) depends on the number of sample points N and the variance of the integrand. The variance $\sigma^2(f)$ of a function f is defined by

$$\sigma^2(f) = \int_{\overline{I}^s} \Big(f(\mathbf{u}) - E(f) \Big)^2 d\mathbf{u},$$

where $E(f) = \int_{\overline{I}} f(\mathbf{u}) d\mathbf{u}$ denotes the expected value of the integral of f. An estimate of the expected error in the Monte Carlo approximation is obtained by averaging the error using every possible random vector $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ to yield

$$\int_{\overline{I}} \int_{\overline{I}} \cdots \int_{\overline{I}^s} \left(\frac{1}{N} \sum_{n=1}^{N} f(\mathbf{x}_n) - E(f) \right)^2 d\mathbf{x}_1 d\mathbf{x}_2 \cdots d\mathbf{x}_N = \frac{\sigma^2(f)}{N}. \tag{3.2}$$

The square root of (3.2) is referred to as the standard error and is equal to $\sigma(f)/\sqrt{N}$, where $\sigma(f)$ is the standard deviation of f. If the standard deviation is bounded, the central limit theorem [47] yields a stronger form for the error bound

$$\lim_{N\to\infty} \text{Prob}\left[\frac{c_1\sigma(f)}{\sqrt{N}} \leq \frac{1}{N}\sum_{n=1}^{N} f(\mathbf{x}_n) - E(f) \leq \frac{c_2\sigma(f)}{\sqrt{N}}\right] = \frac{1}{\sqrt{2\pi}}\int_{c_1}^{c_2} e^{-t^2/2}dt. \quad (3.3)$$

Therefore, the expected convergence rate of Monte Carlo integration is $\mathcal{O}\left(N^{-1/2}\right)$, where N is the number of samples. It is important to note that the error bound (3.3) is independent of the dimension of the integral.

Often for practical applications of Monte Carlo integration, one is interested in the relative error of the approximation; that is, the magnitude of the error normalized by the expected value. Using the result in (3.3), the relative error is bound, with 95% certainty, [2] by the following inequality,

$$\frac{\left|\frac{1}{N}\sum_{n=1}^{N} f(\mathbf{x}_n) - E(f)\right|}{E(f)} \leq \frac{1.4}{\sqrt{N}}\left(\frac{\sigma(f)}{E(f)}\right). \quad (3.4)$$

Note that the relative error is directly proportional to the ratio of the standard deviation to the expected value $\sigma(f)/E(f)$ in addition to the $\mathcal{O}\left(N^{-1/2}\right)$ dependence on sample size. Suppose, for example, the standard deviation of some function is ten times greater than the expected value of its integral, *i.e.* $\sigma(f)/E(f) = 10$. In order to approximate this integral to within 1% of its true value (with a 95% confidence interval) using the Monte Carlo method, one is required to average $N = 2 \cdot 10^6$ independent samples of the function. Hence, for low speed, non-equilibrium gas flows, where the standard deviation of the velocity distribution function can be more than 1000 times greater the bulk flow velocity, the number samples required for an accurate DSMC simulation can extend beyond the trillions.

[2] The actual probability the relative error is within the prescribed bound is $\text{erf}(1.4) \approx 0.95228512$.

3.1.2 Quasi-Monte Carlo Integration Error

Quasi-Monte Carlo (QMC) integration is carefully designed to improve the convergence rate of Monte Carlo by substituting the random variates in (3.1) with more evenly distributed samples. The name QMC applies to any integration approximation that uses equally weighted samples like Monte Carlo (3.1), but offers a near linear theoretical convergence rate $\mathcal{O}\left(N^{-1}(\log N)^{s-1}\right)$. The near linear error convergence rate of QMC represents a significant improvement to the $\mathcal{O}\left(N^{-1/2}\right)$ convergence rate of Monte Carlo. For the probabilistic simulation of low speed non-equilibrium gas flows, achieving near linear convergence with QMC has the potential to improve the computation time required for such flows by orders of magnitude over traditional DSMC. In order to understand how it is possible for QMC integration to obtain this dramatic increase in performance, one must begin with the cornerstone of QMC theory – the Koksma-Hlawka inequality (see [127] p. 18 for a proof).

Theorem 3.1 (Koksma-Hlawka inequality) *If f has bounded variation $V_{HK}(f)$ on I^s in the sense of Hardy and Krause, then, for any point set $P = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \in I^s$, one has*

$$\left| \frac{1}{N} \sum_{i=1}^{N} f(\mathbf{x}_i) - \int_{[0,1]^s} f(\mathbf{u}) d\mathbf{u} \right| \leq V_{HK}(f) D_N^*(P), \tag{3.5}$$

where $D_N^*(P)$ denotes the star discrepancy of the point set P.

Koksma [81] first proved the result (3.5) in one dimension, and Hlawka [66] extended the result to the multi-dimensional case. An explanation and proof of the one dimensional result of Koksma is given in [127] (see Theorem 2.9). The Koksma-Hlawka inequality provides an error bound to any integral approximation using equally weighted samples. Fundamentally, the Koksma-Hlawka inequality separates the integral approximation error into two components; the contribution due to the

integrand behavior, *i.e.* variation $V_{HK}(f)$; and the contribution due to the choice of sample points, *i.e.* discrepancy $D_N^*(P)$. In contrast to the probabilistic error associated with Monte Carlo (3.3), the Koksma-Hlawka inequality is entirely deterministic for a given function and point set.

The goal of QMC integration is to improve on Monte Carlo error convergence for any integrand with bounded variation. In order to accomplish this, QMC strives to use point sets for sampling the integrand with a theoretical star discrepancy lower than the average random sequence used for Monte Carlo. The star discrepancy of a random sequence has an $\mathcal{O}\left(N^{-1/2}(\log\log N)^{1/2}\right)$ convergence rate (expected) [47]. Given the extremely slow growth of the $(\log\log N)$ term, the error bound from the Koksma-Hlawka inequality (3.5) is roughly the same order as obtained from the Central Limit Theorem $\mathcal{O}\left(N^{-1/2}\right)$. Rather than using random variates, QMC integration generates its samples from point sets that have a star discrepancy with an asymptotic convergence of $\mathcal{O}\left(N^{-1}(\log N)^{s-1}\right)$, where s is the dimension of the integrand. Sequences and point sets that achieve this near linear convergence are referred to as low-discrepancy sequences and low-discrepancy points sets. As a consequence of the Koksma-Hlawka inequality (3.5), the theoretical error convergence of QMC integration using a low-discrepancy sequence is $\mathcal{O}\left(N^{-1}(\log N)^{s-1}\right)$, which is superior to Monte Carlo method. In order to understand the construction and implementation of the QMC method, the concepts of discrepancy and variation are discussed in Sections 3.2 and 3.3 respectively.

3.2 Discrepancy

Informally, the discrepancy of a point set P is a measurement of how evenly distributed, or balanced, the points of P are within a specified domain D. For an

"even distribution" of points, there should be no regions of the domain that have an appreciably higher or lower density of points relative to the rest of the domain. This notion is quantified by the discrepancy as the average difference between the fraction of points from P in any subregion $d \subseteq D$ and the volume ratio of d to D. Hence, a point set with a smaller discrepancy yields a more balanced distribution of points throughout the domain. A set of points is uniformly distributed if every subregion $d \subseteq D$ contains a fraction of the point set equal to the volume ratio d to D. In this case, which can only occur for an infinite point set, the discrepancy of the uniformly distributed point set is zero. Intuitively, a more balanced distribution of sample points throughout the integration domain should yield a more accurate approximation of the integral in (3.5). It is precisely this connection between the distribution of sample points and the integration error that the Koksma-Hlawka inequality captures with the discrepancy measure. In regards to general domains, for the reasons mentioned earlier in Section 3.1, all analysis in this investigation is limited to the unit hypercube.

Formally, the definition of the discrepancy of a point set requires a family of sets of the unit hypercube and a norm over family of sets to be specified. For a point set $P \in \overline{I}^s$ with N points, the discrepancy $D_N(P; \mathcal{F})$ over the family of sets $\mathcal{F}$ is defined as

$$D_N(P; \mathcal{F}) = \underset{B \in \mathcal{F}}{\text{norm}} \left(\frac{A(P; B)}{N} - \lambda_s(B) \right), \tag{3.6}$$

where $A(P; B)$ is the counting function that denotes the number of points from P that are contained within the set B. Note that every set $B \in \mathcal{F}$ is contained within $\overline{I}^s$.

There are four basic families of sets that are commonly used in the definition of discrepancy. The first family of sets is $\mathcal{J}^*$, and consists of all the half-open intervals

on $\overline{I}^s$ with one vertex at the origin; that is

$$\mathcal{J}^* = \left\{ \prod_{i=1}^{s} [0, a_i) \ : \ 0 \leq a_i \leq 1 \text{ and } 1 \leq i \leq s \right\}. \tag{3.7}$$

The family $\mathcal{J}^*$ is used to define the star discrepancy found in the Koksma-Hlawka inequality (3.5). The second family of sets is $\mathcal{J}$, which is a generalization of $\mathcal{J}^*$, and consists of all the half-open intervals on $\overline{I}^s$; that is

$$\mathcal{J} = \left\{ \prod_{i=1}^{s} [a_i, b_i) \ : \ 0 \leq a_i < b_i \leq 1 \text{ and } 1 \leq i \leq s \right\}. \tag{3.8}$$

The family $\mathcal{J}$ is used to define the extreme discrepancy. The third family of sets is $\mathcal{H}$, and consists of all the half-spaces obtained by a hyperplane intersecting $\overline{I}^s$. That is, $\mathcal{H}$ is the set of point $\mathbf{x} \in \overline{I}^s$ that satisfies

$$\mathbf{n} \cdot (\mathbf{x} - \mathbf{x}_0) \geq 0, \tag{3.9}$$

for all unit vectors $\mathbf{n} \in \mathbb{R}^s$, and all $\mathbf{x}_0 \in \overline{I}^s$. The family $\mathcal{H}$ is used to define a type of isotropic discrepancy that appears in the study of 3D computer graphics rendering. The fourth family of sets is $\mathcal{C}$, and consists of all the convex polytopes contained within $\overline{I}^s$. The family $\mathcal{C}$ is used to define the general isotropic discrepancy encountered in mathematics literature. However, the general isotropic discrepancy defined by $\mathcal{C}$ is nearly impossible to compute in practice; as such, no explicit form of the family $\mathcal{C}$ is provided here.

In order to complete the definition in (3.6), it is necessary to discuss the concept of a norm over a family of sets. The most common discrepancy norm found in the discussion of QMC integration is the norm used for the star discrepancy, which is the L_∞ norm. The L_∞ norm over a family of sets is defined in analogous manner as the L_∞ norm of a vector. Specifically, the L_∞ discrepancy of the point set P over a

general family of sets $\mathcal{F}$ is given by

$$D_N(P;\mathcal{F}) = \sup_{B\in\mathcal{F}} \left| \frac{A(P;B)}{N} - \lambda_s(B) \right|. \tag{3.10}$$

The L_∞ discrepancy is the measure of the maximum deviation of the point density from the average density of points throughout the domain. When the point set is used to sample a function for an integral approximation, the L_∞ discrepancy is necessary to establish the deterministic upper bound on the error in the Koksma-Hlawka inequality (3.5).

Another discrepancy norm encountered less frequently is the L_2 norm, which again is similar to its vector space analogue. For families of sets where the member sets are easily defined by a few independent variables, an L_2 norm of the argument in (3.6) can be obtained by integrating over all the member sets. Consider, for example, the family of sets $\mathcal{J}^*$. Define each set $B \in \mathcal{J}^*$ by a single vector $\mathbf{x} \in \overline{I}^s$ such that $B(\mathbf{x}) = \prod_{i=1}^{s}[0, x_i)$. In this case, let T_N^* denote the L_2 discrepancy of the point set P over $\mathcal{J}^*$; hence,

$$T_N^*(P) = \left(\int_{\mathbf{x}\in\overline{I}^s} \left(\frac{A\big(P;B(\mathbf{x})\big)}{N} - \prod_{i=1}^{s} x_i \right)^2 d\mathbf{x} \right)^{1/2}. \tag{3.11}$$

The advantage of the L_2 discrepancy is that it can be calculated directly without a costly search for the supremum value, unlike the L_∞ norm. In general, either advanced orthogonal range searching structures or exhaustive searches are required to find the exact L_∞ discrepancy of a general point set, which are discussed in greater detail in the following section. More recent work, by Hickernell [62], extends the concept of the L_2 norm over a family of sets to the general L_p norm in order to create an entire class of L_p discrepancies.

3.2.1 Calculation of the star discrepancy D_N^*

The star discrepancy D_N^* of a point set P is the L_∞ norm over the family of sets $\mathcal{J}^*$ given in (3.7). Hence, $D_N^*(P)$ is defined as

$$D_N^*(P) = \sup_{B \in \mathcal{J}^*} \left| \frac{A(P;B)}{N} - \lambda_s(B) \right|. \tag{3.12}$$

The star discrepancy D_N^* is the most common discrepancy measure found in the QMC literature because of its role in the Koksma-Hlawka inequality (3.5). Moreover, the asymptotic convergence of most low-discrepancy sequences is established in terms of their star discrepancy. For these reasons, the star discrepancy is given the greatest attention in this investigation with regards to the calculation of discrepancy.

For a one dimensional sequence $P = (x_1, \ldots, x_N)$, ordered such that $0 \leq x_1 \leq x_2 \leq \cdots \leq x_n \leq 1$, the star discrepancy of P is calculated from

$$D_N^*(P) = \frac{1}{2N} + \max_{1 \leq n \leq N} \left| x_n - \frac{2n-1}{2N} \right|. \tag{3.13}$$

A derivation of (3.13) is found in [127] (see Theorem 2.6). The calculation in (3.13) requires N evaluations of the absolute value term in order to obtain the maximum. However, this is under the assumption of an ordered point set, which is generally not the case. The cost of sorting a point set is, on average, $\mathcal{O}(N \log N)$ using any of the popular techniques like Quicksort or Heapsort [31, 79]. Therefore, the cost of calculating the one dimensional star discrepancy $D_N^*(P)$ is also $\mathcal{O}(N \log N)$.

For a multi-dimensional sequence, the process of calculating the star discrepancy becomes decidedly more complex. The presence of the discontinuous counting function $A(P;B)$ prevents popular methods employed for well-behaved functions, such as the conjugate gradient method [147], from being used to find the supremum value in (3.12). Left with only a brute force search of an infinite set of intervals, it is necessary to reduce $\mathcal{J}^*$ to a finite set of candidate intervals $\mathcal{B}^*$. Restricting the candidate

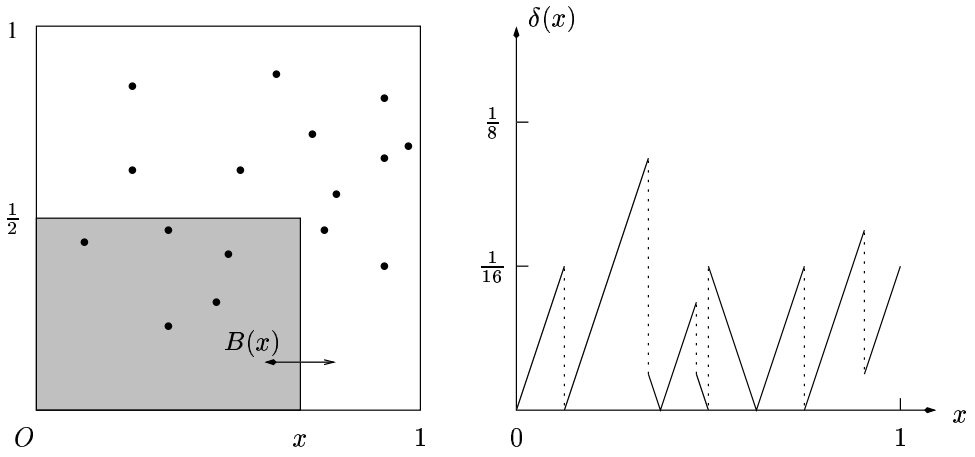

Figure 3.1: Illustration of the possible supremum values in the calculation of the star discrepancy $D_N^*(P)$ of a two dimensional point set P.

intervals to only those that produce local extrema for the absolute value function in (3.12), reduces the number of intervals that must be checked to find the supremum value. Fortunately, for a finite point set, the number of candidate intervals in $\mathcal{B}^*$ is also finite.

In order to illustrate how the reduction in candidate intervals is achieved, consider the random two dimensional sequence given in Figure 3.1. For the set of intervals $B(x)$ defined as

$$B(x) = [0, x) \times [0, 1/2), \text{ for } 0 \leq x \leq 1,$$

the absolute value function from (3.12),

$$\delta(x) = \left| \frac{A\big(P; B(x)\big)}{N} - \lambda_s\big(B(x)\big) \right|,$$

is also plotted in Figure 3.1. Note that the local extrema of $\delta(x)$ occur whenever x is equal to the location of a point in P with a y coordinate less than $1/2$ or when $x = 1$. Specifically, these locations correspond to intervals with a point in P on the $x-$ boundary of the interval. Therefore, to find the supremum value over the sets $B(x)$, one only needs to evaluate the function $\delta(x)$ at seven values of x (technically 13 locations– as both sides of the discontinuous jump should be checked).

In fact, for any multi-dimensional point set P, the set of candidate intervals for the supremum search can be reduced to only intervals that possess points in P or 1 on each boundary. To state this in a slightly more general form, let $P = (\mathbf{x_1}, \ldots, \mathbf{x_N})$ represent an s-dimensional set of points, with each point for $1 \leq n \leq N$ defined as $\mathbf{x_n} = (x_n^{(1)}, \ldots, x_n^{(s)})$. Next let P_i represent the set of the i^{th} coordinates from each point $\mathbf{x_n} \in P$, that is $P_i = (x_1^{(i)}, \ldots, x_N^{(i)})$. The set of candidate intervals $\mathcal{B}^*$ is then defined by

$$\mathcal{B}^* = \left\{ \prod_{i=1}^{s} [0, x_i) \text{ and } \prod_{i=1}^{s} [0, x_i] \ : \ x_i \in P_i \cup 1, \ 1 \leq i \leq s \right\}. \tag{3.14}$$

Note that both the half-closed and closed intervals must be included in $\mathcal{B}$ for the following reason. Given a point $p \in P_i$, the local maxima can occur in either direction $x_i \to p^-$ and $x_i \to p^+$, as demonstrated by the discontinuous jumps in Figure 3.1. The maximum number of potential candidates for the supremum in (3.12) is now reduced to a finite value equal to $\text{card}(\mathcal{B}) = (2(N+1))^s$. Hence,

$$D_N^*(P) = \sup_{B_c \in \mathcal{B}} \left| \frac{A(P; B_c)}{N} - \lambda_s(B_c) \right|. \tag{3.15}$$

In order to understand why the supremum search in (3.12) can be reduced to the finite set of intervals that have a point in P or 1 on each boundary, consider the following argument. Suppose the supremum value of (3.12) is obtained for the closed interval $B(\mathbf{x}) = \prod_{i=1}^{s} [0, x_i] \in \mathcal{J}^*$ defined by the vector $\mathbf{x} = (x_1, \ldots, x_s)$, and that $B(\mathbf{x}) \notin \mathcal{B}^*$. Then, there exists at least one dimension of $B(\mathbf{x})$ without a point from P on its boundary; that is to say $x_k \notin P_k$ for some k. If such a dimension of the interval $B(\mathbf{x})$ were to exist, one could perturb slightly the interval to change its volume measure without changing the number of points from P that it contains. Thus, there would exist vectors $\mathbf{x}^- = (x_1, \ldots, x_k - \epsilon, \ldots, x_s)$ and $\mathbf{x}^+ = (x_1, \ldots, x_k + \epsilon, \ldots, x_s)$,

defined for $\epsilon > 0$, such that

$$A\big(P; B(\mathbf{x}^-)\big) = A\big(P; B(\mathbf{x})\big) = A\big(P; B(\mathbf{x}^+)\big),$$

and

$$\lambda_s\big(B(\mathbf{x}^-)\big) < \lambda_s\big(B(\mathbf{x})\big) < \lambda_s\big(B(\mathbf{x}^+)\big)$$

Therefore, the absolute value term in (3.12) for either $B(\mathbf{x}^-)$ or $B(\mathbf{x}^+)$ will be greater than the supremum value associated with $B(\mathbf{x})$, which contradicts the initial assumption. Hence, a closed interval must be in $\mathcal{B}^*$ if it is to yield the supremum value for the discrepancy. The same argument can be applied to half-open intervals to conclude that the supremum value in (3.12) must be associated with a candidate interval in $\mathcal{B}^*$.

While the family of sets $\mathcal{B}^*$ contain all the possible candidate intervals for the supremum value in (3.12), not every interval in $\mathcal{B}^*$ is necessarily a valid candidate. Specifically, the definition of $\mathcal{B}^*$ decouples the dimensions of each point in P making it possible to have a candidate interval without a point from P on the boundary. For example, consider the simple set $P = \{(\frac{1}{4}, \frac{1}{3}), (\frac{1}{2}, \frac{2}{3})\}$. The interval $[0, \frac{1}{4}) \times [0, \frac{2}{3})$ is in $\mathcal{B}^*$, but it does not have a point on its $y = \frac{2}{3}$ boundary. In addition, points in P that share a common coordinate value produce duplicate intervals in $\mathcal{B}^*$ that can also be excluded. However, the number of potential candidate intervals in $\mathcal{B}^*$ for the supremum search remains $\mathcal{O}(N^s)$, even after excluding these pathological cases.

In order to determine the computational complexity required for calculating the multi-dimensional star discrepancy, the cost of evaluating each candidate interval in the supremum search of $\mathcal{B}^*$ is needed. For each candidate interval, the absolute value term in (3.15) is calculated to determine its supremum value. The main cost of evaluating the absolute value term is attributed to the counting function $A(P; B_c)$.

Simply counting the number of $x \in P$ that are also $x \in B_c$ will calculate $A(P; B_c)$ in $\mathcal{O}(N)$ steps, yielding an overall calculation cost of $\mathcal{O}(N^{s+1})$ for the star discrepancy. However, it is possible to calculate $A(P; B_c)$ without querying every point in P if the point set is ordered in some manner. More specifically, if the points of P are stored in an s dimensional orthogonal range tree [37], then $A(P; B_c)$ can be calculated in $\mathcal{O}((\log N)^s)$ steps. The one time cost of constructing the orthogonal range tree is $\mathcal{O}(N(\log N)^{s-1})$,[3] yielding an overall calculation cost of $\mathcal{O}(N(\log N)^s)$ for computing the star discrepancy.

The computational savings obtained using the orthogonal range trees for the calculation of $A(P; B_c)$ decreases as the sequence dimension increases. Moreover, as the sequence dimension increases, the data structures associated with orthogonal range trees become increasing more complicated. Regardless of the means used to calculate $A(P; B_c)$, the star discrepancy calculation is always limited by the complexity of the candidate interval set, which is $\mathcal{O}(N^s)$. Thus, for the large sequences (in both dimension and length) needed in this investigation, any exhaustive search of the candidate intervals $B_c \in \mathcal{B}^*$ for the supremum is impractical.

Instead of trying to find the exact value of the star discrepancy $D_N^*(P)$ using an exhaustive search for the supremum in (3.12), one can obtain an estimate $\overline{D}_N^*(P)$ by limiting the supremum search to an even smaller set of intervals $\mathcal{S}$ than the family $\mathcal{B}^*$. That is,

$$D_N^*(P) \approx \overline{D}_N^* = \sup_{B \in \mathcal{S}} \left| \frac{A(P; B)}{N} - \lambda_s(B) \right|. \tag{3.16}$$

Further computational savings can be realized if one defines the candidate intervals in $\mathcal{S}$ independently of the point set P, because the calculation of the counting function is greatly simplified. In this case, the points of P are first sorted into the intervals of

[3]The memory cost of storing the orthogonal range tree is also $\mathcal{O}\left(N(\log N)^{s-1}\right)$.

$\mathcal{S}$ via the Pigeon-Hole sort (or counting sort) [31] in linear time $\mathcal{O}(N)$.[4] Then, with the points sorted, the counting function can be calculated in constant time $\mathcal{O}(1)$ for every candidate interval in $\mathcal{S}$. While the resulting Pigeon-Hole algorithm for the star discrepancy estimate $\overline{D}_N^*$ is rather inelegant, it is efficient with a total computational cost of $\mathcal{O}(N + \text{card}(\mathcal{S}))$. Thus, by choosing the set $\mathcal{S}$ such that $\text{card}(\mathcal{S}) \ll \text{card}(\mathcal{B}^*)$, the performance is much faster than the exact calculation using the supremum search. Furthermore, the Pigeon-Hole algorithm for the star discrepancy estimate $\overline{D}_N^*$ does provide a consistent approximation with a known error bound. It is to be expected that the assured accuracy of the star discrepancy estimate $\overline{D}_N^*$ will decrease when the number of candidate intervals in $\mathcal{S}$ decreases.

For this investigation, the reduced set $\mathcal{S}$ in the Pigeon-Hole algorithm for the star discrepancy estimate $\overline{D}_N^*$ is taken to be the set of equally spaced intervals defined by

$$\mathcal{S} = \left\{ \prod_{i=1}^{s} [0, \tfrac{n_i}{M}) \ : \ 1 \leq n_i \leq M \right\}, \tag{3.17}$$

where M is a constant denoting the number of intervals in each dimension. The number of intervals that must be checked to find the supremum in (3.16) is M^s. Since the intervals are equally spaced, the Pigeon-Hole sort is accomplished in $\mathcal{O}(N)$ time yielding an overall computational cost of $\mathcal{O}(N + M^s)$. The accuracy of the star discrepancy approximation using the definition of $\mathcal{S}$ in (3.17) is found to be

$$0 \leq D_N^*(P) - \overline{D}_N^*(P) \leq 1 - \left(1 - \frac{1}{M}\right)^s \approx \mathcal{O}\left(M^{-1}\right). \tag{3.18}$$

A more elaborate algorithm for approximating the star discrepancy with a reduced set of candidate intervals is considered by Thiémard [175], which allows for variably spaced intervals to improve the accuracy.

[4]The Pigeon-Hole sort is performed in linear time when their exists a continuous map from the sorting range to the sorting intervals (or bins) that can be computed in constant time, *e.g.* intervals of equal size. Otherwise, the sorting interval is determined using $\mathcal{O}(\log N)$ comparisons, which yields an overall computation time of $\mathcal{O}(N \log N)$.

In order to understand its actual implementation, the Pigeon-Hole algorithm for the star discrepancy estimate $\overline{D}_N^*$ is presented in greater detail here. The algorithm can be extended to any number of dimensions, but in the interest of keeping the index notation as simple as possible, only the calculation in two dimensions is considered here. Define the matrices $E, A \in \mathbb{R}^{M\times M}$ to represent the counting function over the following discrete intervals:

$$\begin{aligned} E_{mn} &= A\big(P; [\tfrac{m-1}{M}, \tfrac{m}{M}) \times [\tfrac{n-1}{M}, \tfrac{n}{M})\big) \\ A_{mn} &= A\big(P; [0, \tfrac{m}{M}) \times [0, \tfrac{n}{M})\big). \end{aligned}$$

The matrix E is calculated from the Pigeon-Hole sort of the sequence P, with a computational cost $\mathcal{O}(N)$. The matrix A is determined from the matrix E through the recursive construction given by

$$\begin{aligned} &A_{11} = E_{11}, && \\ &A_{m1} = E_{m1} + A_{m-1,1} && 2 \le m \le M, \\ &A_{1n} = E_{1n} + A_{1,n-1} && 2 \le n \le M, \\ &A_{mn} = E_{mn} + A_{m-1,n} + A_{m,n-1} - A_{m-1,n-1} && 2 \le m, n \le M. \end{aligned} \tag{3.19}$$

The calculation of the matrix A using (3.19) is performed in $\mathcal{O}(M^2)$ time. Once the matrix A is obtained for a given point set P, the approximation to the star discrepancy $\overline{D}_N^*(P)$ is calculated by

$$\overline{D}_N^*(P) = \max_{1\le m,n\le M} \left| \frac{A_{mn}}{N} - \frac{mn}{M^2} \right|. \tag{3.20}$$

The maximum value of (3.20) is tracked while the matrix A is constructed; thus, the total cost for estimating the star discrepancy $\overline{D}_N^*(P)$ is $\mathcal{O}(N + M^2)$. Furthermore, from the result in (3.18), the error in the star discrepancy approximation $\overline{D}_N^*(P)$ is

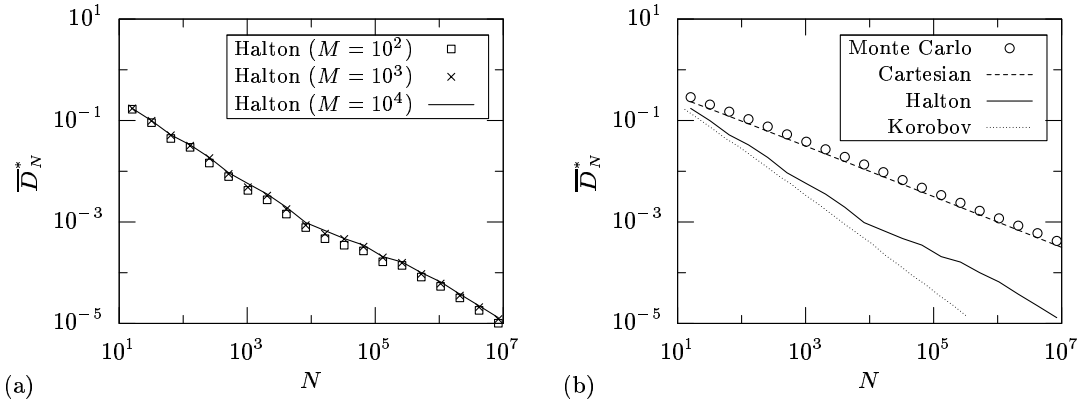

Figure 3.2: Convergence of the approximate star-discrepancy $\overline{D}^*_N$ in two dimensions: (a) varying the numbers of sorting bins M for the Halton sequence; and (b) comparing several different types of point sets.

at most given by

$$D^*_N(P) - \overline{D}^*_N(P) < \frac{2}{M} - \frac{1}{M^2}. \tag{3.21}$$

Example calculations using the preceding algorithm to approximate the star discrepancy (3.20) of the low-discrepancy Halton sequence [56] are presented in Figure 3.2(a). In particular, three different values of M, the number of discrete intervals in each dimension, are considered for the Halton sequence. Using the result in (3.21), the maximum possible error associated with approximate star discrepancy $\overline{D}^*_N(P)$ is roughly 0.02, 0.002, and 0.0002 when $M = 10^2$, 10^3, and 10^4, respectively. However, in Figure 3.2(a), there does not appear to be a significant difference between the star discrepancy approximations despite the maximum possible error varying by two orders of magnitude. It is important to note that the upper bound in (3.21) corresponds to the rather improbable pathological case when all the points are distributed arbitrarily close to the domain edges at $x = 1$ and $y = 1$. The lack of sensitivity to number of discrete intervals M is due in large part to the fact that, as a low-discrepancy sequence, the construction of the Halton sequence attempts to maintain a near constant point density throughout the domain. Hence, for any size interval,

the fraction of points contained within the interval should scale appropriately.

Further examples of the approximate star discrepancy calculation (3.20) are given in Figure 3.2(b) to demonstrate the convergence of some common point sets. As expected, the star discrepancy of the random sequence used in the Monte Carlo converges roughly as $\mathcal{O}\left(N^{-1/2}\right)$ for N samples. The low-discrepancy Halton sequence has a star discrepancy that is much better than the random sequence for all sequence lengths presented. The star discrepancy of the Halton sequence has an initial convergence rate that is near linear; however, for $N > 10^4$, the convergence rate is less than linear, yet still better than random. It is important to note that the less than theoretical convergence observed for the Halton sequence does not preclude it from being a low-discrepancy point set. The Halton sequence receives a low-discrepancy classification because it can be proven mathematically that the star discrepancy converges asymptotically $D_N^*(P) \approx \mathcal{O}\left(N^{-1}(\log N)^{s-1}\right)$ as N tends to infinity. Thus, the results in Figure 3.2 are only indicative of the performance of the Halton sequence for these set sizes.

Another example of a low-discrepancy point set is the optimum integration lattice obtained from Korobov's method of good lattice points [82, 83]. In Figure 3.2, the optimum lattice achieves an even faster star discrepancy convergence than the Halton sequence. In general, optimum integration lattices typically have a lower discrepancy than low-discrepancy sequences for the same number of points; the reasons for this are presented later in Section 3.4. The final example to note in Figure 3.2 is the set of equally spaced Cartesian grid points. This is the same set of sample points used by the open Newton-Cotes formulas for integral approximations [20]. The Newton-Cotes formulas can be extremely powerful for integrals in one dimension when different sample weights are selected to achieve a high order accuracy.

However, methods that sample an integrand at Cartesian grid points have an error convergence rate that decreases significantly with increasing dimension. This property shared by the Newton-Cotes formulas and Gaussian quadrature is referred to as "the curse of dimensionality [127]." For the open Newton-Cotes formulas, the poor error convergence in large dimensions is indicated by the asymptotic convergence of the star discrepancy, which is $\mathcal{O}\left(N^{1/s}\right)$ (see Theorem 3.14 in [127]). In Figure 3.2, the observed convergence rate of the star discrepancy of the two dimensional Cartesian grid points is consistent with theory (*i.e.* $\mathcal{O}\left(N^{-1/2}\right)$), and appears the same as a random sequence.

3.2.2 Calculation of the extreme discrepancy D_N

The extreme discrepancy D_N of a point set P is the L_∞ norm over the family of sets $\mathcal{J}$ given in (3.8). Hence, $D_N(P)$ is defined as

$$D_N(P) = \sup_{B\in\mathcal{J}} \left| \frac{A(P;B)}{N} - \lambda_s(B) \right|. \tag{3.22}$$

An upper and lower bound on the extreme discrepancy is established in terms of the star discrepancy such that

$$D_N^* \leq D_N \leq 2^s D_N^*. \tag{3.23}$$

The family of sets $\mathcal{J}^*$, that the star discrepancy is based upon, is a strict subset of the family of sets $\mathcal{J}$; hence, the star discrepancy serves as a lower bound for the extreme discrepancy in (3.23). By constructing the counting function and volume measure of a general interval in $\mathcal{J}$ from a linear combination of intervals in $\mathcal{J}^*$, the upper bound on the extreme discrepancy (3.23) is obtained in [85].

For a one dimensional sequence $P = (x_1, \ldots, x_N)$, ordered such that $0 \leq x_1 \leq x_2 \leq \cdots \leq x_n \leq 1$, the extreme discrepancy of P is calculated from

$$D_N(P) = \frac{1}{N} + \max_{1\leq n\leq N}\left(\frac{n}{N} - x_n\right) - \min_{1\leq n\leq N}\left(\frac{n}{N} - x_n\right).$$

A proof of the preceding result is given in [127]. As with the star discrepancy calculation (3.13), the asymptotic computational cost for the extreme discrepancy of a one dimensional sequence is $\mathcal{O}(N \log N)$; because the calculation is limited by the sorting of the point set.

The exact calculation of the extreme discrepancy for a multi-dimensional sequence follows the same procedure as the star discrepancy calculation described in Section 3.2.1. The key difference is that the intervals in $\mathcal{J}$ do not require a vertex at the origin. Using the same argument as in the star discrepancy calculation, each dimension of a candidate interval, for the supremum search in (3.22), must have points in P or 0 or 1 for a boundary. Hence, the number of candidate intervals in the exact calculation of the multi-dimensional extreme discrepancy increases to $\mathcal{O}(N^{2s})$. The cost of evaluating the counting function in (3.22) is the same as in the star discrepancy calculation; that is, $\mathcal{O}((\log N)^s)$ if the points in P are stored in an orthogonal range tree. Thus, the total cost of calculating the exact extreme discrepancy is $\mathcal{O}((N^2 \log N)^s)$. Similar reasoning can be applied to the approximate calculation of the extreme discrepancy to yield an $\mathcal{O}(N + M^{2s})$ computational cost, using M equally spaced intervals in each dimension for the supremum search.

3.2.3 Calculation of the quadratic mean discrepancies T_N^* and T_N

The quadratic mean star discrepancy T_N^* of a point set P is the L_2 norm over the family of sets $\mathcal{J}^*$ given in (3.7). The explicit definition of $T_N^*(P)$ was given previously in (3.11). Given the analogue between D_N^* and the L_∞ norm, it is natural that the star discrepancy is an upper bound on the quadratic-mean star discrepancy; that is

$$C_s\big(D_N^*(P)\big)^{\frac{s+2}{2}} \leq T_N^*(P) \leq D_N^*(P),$$

where the lower bound on the quadratic-mean star discrepancy is proven by Niederreiter in [124]. Note that C_s is a function that depends only on the dimension of the sequence s. It is possible to establish an integration error bound similar to the Koksma-Hlawka inequality (3.5) using T_N^* instead of D_N^*, see Zaremba's proof in [192] for functions with continuous mixed partial derivatives. Also, an even more direct relationship between the quadratic-mean discrepancy and the expected integration error is established by Woźniakowski in [190]. The results in [190] offer a much more optimistic error bound than the Koksma-Hlawka inequality for a certain class of functions that are not of bounded variation in the sense of Hardy and Krause.

In [183], Warnock establishes that $T_N^*(P)$ can be calculated directly from

$$(T_N^*)^2 = 3^{-s} + \frac{1}{N^2}\sum_{m=1}^{N}\sum_{n=1}^{N}\prod_{i=1}^{s}(1-\max(x_{m,i},x_{n,i})) - \frac{2^{-s+1}}{N}\sum_{n=1}^{N}\prod_{i=1}^{s}(1-x_{n,i}^2). \quad (3.24)$$

The double sum in the calculation (3.24) implies that the asymptotic computation cost is $\mathcal{O}(N^2)$ with an implied coefficient that increases linearly with the dimension s. For a multi-dimensional sequence, this is a vast improvement over the calculation of the star discrepancy D_N^* in Section 3.2.1. Moreover, the calculation is easier to implement than the star-discrepancy calculation because there is no need for sorting the point set. As outlined by Heinrich in [59], some calculation of the double sum can be avoided, if the ordering of the points is exploited, to yield a computation cost of $\mathcal{O}(N(\log N)^s)$.

The quadratic mean extreme discrepancy T_N of a point set P is the L_2 norm over the family of sets $\mathcal{J}$ given in (3.8). Hence, $T_N(P)$ is defined as

$$T_N(P) = \left(\int\limits_{\substack{\mathbf{x},\mathbf{y}\in\overline{I}^s \\ x_i<y_i,\, 1\le i\le s}} \left(\frac{A\big(P;B(\mathbf{x},\mathbf{y})\big)}{N} - \prod_{i=1}^{s}(y_i-x_i) \right)^2 d\mathbf{x} \right)^{1/2}.$$

Using the same approach as in [183], Morokoff in [116] establishes the following direct

calculation for $T_N(P)$:

$$\begin{aligned} T_N^2 &= 12^{-s} + \frac{1}{N^2}\sum_{m=1}^{N}\sum_{n=1}^{N}\prod_{i=1}^{s}\min(x_{m,i}, x_{n,i})\cdot(1-\max(x_{m,i}, x_{n,i})) \\ &\quad -\frac{2^{-s+1}}{N}\sum_{n=1}^{N}\prod_{i=1}^{s}x_{n,i}(1-x_{n,i}). \end{aligned}$$

As before, the double sum in the preceding calculation implies that the asymptotic computation cost of $T_N(P)$ is $\mathcal{O}(N^2)$ with an implied coefficient that increases linearly with the dimension s. The quadratic mean extreme discrepancy does not generally appear in the QMC literature because no direct connection between the integration error and T_N has been established. Although, it is conjectured in [116] that $T_N < T_N^*$.

3.2.4 Calculation of the isotropic discrepancies H_N and J_N

The half-plane discrepancy H_N of a point set P is the L_∞ norm over the family of sets $\mathcal{H}$ given in (3.9). Hence, $H_N(P)$ is defined as

$$H_N(P) = \sup_{B\in\mathcal{H}}\left|\frac{A(P;B)}{N} - \lambda_s(B)\right|. \tag{3.25}$$

The half-plane discrepancy H_N is a useful error measure that occurs in some applications of ray tracing used in 3-D computer graphics rendering. Specifically, when two 3-D objects overlap on the computer screen, pixels on the visual boundary between the objects are often sampled with several rays to approximate the display color. If a boundary pixel is viewed as a unit square, then the half-plane discrepancy $H_N(P)$ is the maximum possible color error produced when sampling the pixel from rays originating at the points in P.

In the context of ray tracing techniques for 3-D computer graphics rendering, a very in-depth description is presented in [37] for the algorithm to calculate the half-plane discrepancy in two dimensions. Similar to the supremum search for the

star and extreme discrepancies in Sections 3.2.1 and 3.2.2, one must consider all the possible candidate half-planes in $\mathcal{H}$ that could produce the supremum value in (3.25). It turns out, there are two types of half-planes that can produce the necessary local extrema of (3.25); hence, one can restrict his/her supremum search to only them. Type 1 half-planes intersect a single point in P and a vertex of the unit square $\overline{I}^2$, and half-planes intersecting a single point in P at such an angle as to maximize or minimize the volume measure $\lambda_s(B)$ in (3.25). The number of candidate Type 1 half-planes is $\mathcal{O}(N)$ because the number of local extrema associated with each single point is finite. Type 2 half-planes intersect any two points in P yielding $\mathcal{O}(N^2)$ potential candidate half-planes.

If one proceeds to calculate the absolute value in (3.25) using an exhaustive query to determine the membership of the points P in the candidate half-plane, the resulting algorithm to calculate H_N in two dimensions requires $\mathcal{O}(N^3)$ steps. However, the algorithm presented in [37] is able to calculate the half-plane discrepancy in $\mathcal{O}(N^2)$ steps, which is optimal in the sense that it is the same order as the computational complexity of the problem. The algorithm in [37] achieves this improvement through the clever use of data structures to exploit the duality transform between the point set P and an arrangement of lines in 2-space. For calculating the half-plane discrepancy H_N in s dimensions, the computational complexity, *i.e.* the number of candidate half-planes for the supremum search, is $\mathcal{O}(N^s)$. Regardless of the algorithm actually used to find the supremum value in (3.25), the computational cost must be at least as great as the computational complexity. Therefore, the cost to calculate the half-plane discrepancy suffers from the same polynomial growth with dimension that plagues the algorithms for the star and extreme discrepancies.

The isotropic discrepancy J_N of a point set P is the L_∞ norm over the family $\mathcal{C}$

of all convex subsets of $\overline{I}^s$. Hence, $J_N(P)$ is defined as

$$J_N(P) = \sup_{B \in \mathcal{C}} \left| \frac{A(P;B)}{N} - \lambda_s(B) \right|. \tag{3.26}$$

Niederreiter and Wills [128][5] establish an upper and lower bound on the isotropic discrepancy J_N in terms of the extreme discrepancy to yield

$$D_N(P) \leq J_N(P) \leq 4sD_N(P)^{1/s}. \tag{3.27}$$

The relatively weak convergence of the upper bound in (3.27) implies that it is possible to have poor error convergence $\mathcal{O}\left(N^{-1/s}\right)$ when using QMC integration on some discontinuous functions. Unfortunately, the exponent $1/s$ in the upper bound of (3.27) that indicates the potential for poor convergence can not be improved upon, as Zaremba [193] demonstrates by example.

In an effort to understand the connection between the isotropic discrepancy $J_N(P)$ and the approximation of a discontinuous integral, consider the following example of a function comprised entirely of convex discontinuities not aligned with the principle axes. Let $F(\mathbf{x})$ denote the composition of M discontinuous functions given by

$$F(\mathbf{x}) = \sum_{i=1}^{M} \alpha_i \phi_{C_i}(\mathbf{x}), \tag{3.28}$$

where α_i is a finite non-zero constant, and ϕ_{C_i} is the characteristic function of the convex set $C_i \in \mathcal{C}$ for $1 \leq i \leq M$. The characteristic function $\phi_C(\mathbf{x})$, with $\mathbf{x} \in \overline{I}^s$ is defined by

$$\phi_C(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in C \\ 0 & \text{otherwise.} \end{cases}$$

Since the discontinuities are not aligned with the principle axes, the function in (3.28) is not of bounded variation in the sense of Hardy and Krause (the concept of

[5]For an English account their results refer to [85, 127]. The result in [85] contains a more detailed analysis but of a slightly weaker result than [128]

variation will be discussed in greater detail in Section 3.3). As such, the Koksma-Hlawka inequality (3.5) can not be used to provide information regarding the error convergence rate of the approximation of the integral of F.

While the Koksma-Hlawka inequality can not be used in this case, an alternate error bound can be established using the isotropic discrepancy defined in (3.26). Starting with the basic error definition, the contribution of each discontinuous characteristic function can be separated by the triangle inequality to yield

$$\begin{aligned}\left|\frac{1}{N}\sum_{n=1}^{N}F(\mathbf{x}_n)-\int_{\bar{I}^s}F(\mathbf{u})d\mathbf{u}\right| &= \left|\sum_{i=1}^{M}\alpha_i\left(\frac{1}{N}\sum_{n=1}^{N}\phi_{C_i}(\mathbf{x}_n)-\int_{\bar{I}^s}\phi_{C_i}(\mathbf{u})du\right)\right| \\ &\leq \sum_{i=1}^{M}\left|\alpha_i\left(\frac{1}{N}\sum_{n=1}^{N}\phi_{C_i}(\mathbf{x}_n)-\int_{\bar{I}^s}\phi_{C_i}(\mathbf{u})du\right)\right| \qquad (3.29)\end{aligned}$$

Note that ϕ_{C_i} is a simple characteristic function for the convex set C_i; and that the terms inside the parentheses in (3.29) are related to the terms from the discrepancy definition via

$$\frac{1}{N}\sum_{n=1}^{N}\phi_{C_i}=\frac{A(P;C_i)}{N},$$

and

$$\int_{\bar{I}^s}\phi_{C_i}(\mathbf{u})du=\lambda_s(C_i).$$

Hence, applying the Hölder inequality to (3.29) yields

$$\left|\frac{1}{N}\sum_{n=1}^{N}F(\mathbf{x}_n)-\int_{\bar{I}^s}F(\mathbf{u})d\mathbf{u}\right|\leq\sum_{i=1}^{M}|\alpha_i|\left|\frac{A(P;C_i)}{N}-\lambda_s(C_i)\right|. \qquad (3.30)$$

Since $C_i\in\mathcal{C}$ for $1\leq i\leq M$, the definition of isotropic discrepancy in (3.26) implies that

$$\left|\frac{A(P;C_i)}{N}-\lambda_s(C_i)\right|\leq\sup_{B\in\mathcal{C}}\left|\frac{A(P;B)}{N}-\lambda_s(B)\right|=J_N(P),$$

which further simplifies the error bound in (3.30) to yield

$$\left|\frac{1}{N}\sum_{n=1}^{N}F(\mathbf{x}_n)-\int_{\bar{I}^s}F(\mathbf{u})d\mathbf{u}\right|\leq J_N(P)\sum_{i=1}^{M}|\alpha_i|. \qquad (3.31)$$

It is important to remember that the result in (3.31) is only an upper bound on the integration error. The use of the triangle and Hölder inequalities along with the supremum value associated with the isotropic discrepancy indicates the bound may not be very tight. However, it does imply that there is a potential for the QMC method to have poor error convergence, possibly as slow as $\mathcal{O}\left(N^{-1/s}\right)$ when the integrand contains discontinuities not aligned with the principle axes. Furthermore, (3.31) suggests that increasing the number and size of the discontinuities of a function will have a negative impact on the accuracy of its integral approximation. Even if a function $F(\mathbf{x}) = f(\mathbf{x}) + \alpha\phi_C(\mathbf{x})$ is composed of a continuous function f and a single convex discontinuity ϕ_C, the potential for poor QMC convergence still exists. The preceding analysis can be repeated in this case to yield the following similar result:

$$\left| \frac{1}{N} \sum_{n=1}^{N} F(\mathbf{x}_n) - \int_{\bar{I}^s} F(\mathbf{u}) d\mathbf{u} \right| \leq D_N^*(P) V_{HK}(f) + \alpha J_N(P).$$

In practice, the observed error convergence of the QMC integral approximation for a function with discontinuities not aligned with the primary axes is better than $\mathcal{O}\left(N^{-1/s}\right)$ (see [115, 117, 120, 146] for examples). However, the presence of such discontinuities does have a significant negative impact on the error; and in general, near linear convergence of the QMC approximation is not obtained. Recall that the result in (3.27) is an upper bound, and only indicates the *potential* for a lower error convergence rate in this case. The bound in (3.27) also implies that the error convergence rate may worsen as the problem dimension increases, which is, in fact, supported by the discontinuous integrals tested in [117]. Despite the degradation in performance with increasing dimension, the error convergence rate of QMC typically remains at least as fast as the Monte Carlo convergence rate; that is, $\mathcal{O}\left(N^{-1/2}\right)$.

Press and Teukolsky [146] suggest that the QMC error convergence rate ap-

proaches $\mathcal{O}\left(N^{-1/2}\right)$ when the surface area of the discontinuities is large relative to the volume of the integration domain. In the neighborhood of a discontinuity, the chance a specific point from a low-discrepancy sequence lands on either side of the discontinuity approaches a random process. Thus, for functions dominated by discontinuous surfaces, the QMC performance approaches that of Monte Carlo. Extending the observation in [146] to a more rigorous framework, Morokoff [117] proves a more optimistic upper bound on the isotropic discrepancy

$$J_N(P) \leq C_s N^{1-s} D_N(P)^{1/(2s-1)}, \tag{3.32}$$

under the assumption of true randomness near the boundary of the discontinuity. Here the coefficient C_s depends only on the dimension s of the problem. The result of Morokoff (3.32) supports the observation that the QMC error rate is at least as fast as the Monte Carlo convergence rate $\mathcal{O}\left(N^{-1/2}\right)$.

The calculation of the isotropic discrepancy $J_N(P)$ is a computationally intractable problem for all but the smallest point sets P. Consider for a moment the calculation of a lower bound to $J_N(P)$ in two dimensions. Instead of finding the supremum in (3.26) over all possible convex sets in $\mathcal{C}$, restrict the supremum search to the family $\mathcal{T}$ consisting of all the triangle sets contained in $\overline{I}^2$. As with the other L_∞ discrepancy calculations, any triangle set that has two points from P on each boundary is a potential candidate for the supremum value taken over $\mathcal{T}$. Thus, the number of candidate triangles in $\mathcal{T}$ is $\mathcal{O}\left(N^6\right)$, where the maximum cost of evaluating the absolute value term in (3.26) for each candidate triangle is $\mathcal{O}\left(N\right)$. The computational complexity of the supremum search for calculating this triangular discrepancy is then between $\mathcal{O}\left(N^6\right)$ and $\mathcal{O}\left(N^7\right)$. An example of the triangular discrepancy calculation using a Monte Carlo approach is given in [60]. In order to calculate the

isotropic discrepancy $J_N(P)$, one must not only consider the supremum over every triangular set, but also over every convex quadrilateral set $\mathcal{O}\left(N^8\right)$, pentagonal set $\mathcal{O}\left(N^{10}\right)$, hexagonal set $\mathcal{O}\left(N^{12}\right)$, and so forth up to an $N-$sided set. Note that this significant amount of computational effort is required for just the two dimensional isotropic discrepancy, which represents considerably more calculation effort than the star and extreme discrepancies in two dimensions.

3.3 Variation

Informally, the variation $V(f)$ of a function f is a measure of the "smoothness" of the function. The continuity of a function and its derivatives affect the magnitude of the variation. If two functions are continuous, the function with a smaller variation measure will typically appear smoother, and have smaller gradients. Geometrically, if an integral is interpreted as the volume contained by the function surface, the variation of the function is a rough measure of its surface area[1]. Consider the following function on the unit interval

$$f_1(x,\tau) = 1 + \cos\left(\tau x^2\right), \tag{3.33}$$

where τ is a positive parameter. The function $f_1(x,\tau)$ is plotted in Figure 3.3(a) for the parameter values $\tau = 10$ and 100. Both functions contain nearly the same volume under the curve; however, the surface area containing the volume is much larger for the $\tau = 100$ case. The function $f_1(x,100)$ is decidedly less smooth than $f_1(x,10)$ because there are 10 times as many function extrema in the $\tau = 100$ case (excluding the end points). The variation measure reflects this change in smoothness and is inexorably connected to the number and magnitude of the local extrema of

[1]Here the terms *volume* and *surface area* refer to their multi-dimensional analogues.

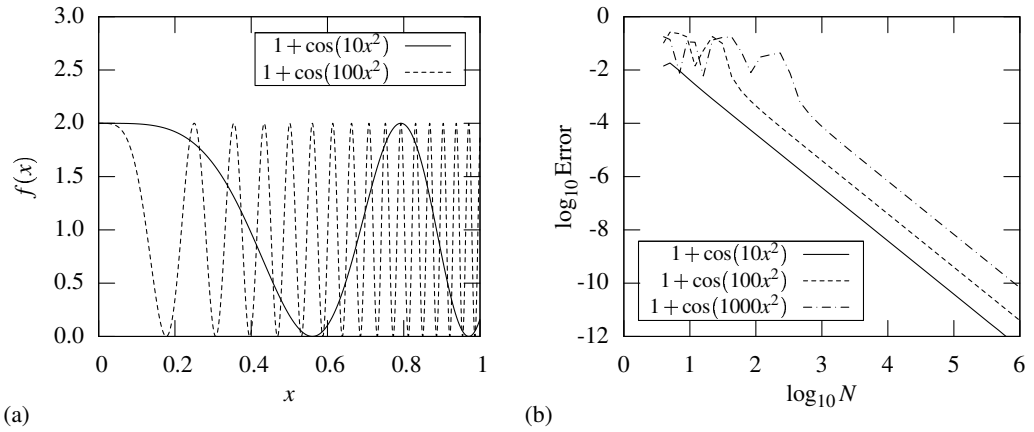

Figure 3.3: Examples of functions with bounded variation: (a) plot of the functions; and (b) integration error convergence.

a function. As such, the variation $V\left(f_1(x, 100)\right)$ is roughly 10 times large than the variation $V\left(f_1(x, 10)\right)$.

In addition to being a measure of "smoothness," a function's variation also serves as a sensitivity measure for the approximation of its integral. It is this role as a measure of sensitivity that is of primary interest to this investigation. Recalling the Koksma-Hlawka inequality (3.5), the error in the integration approximation using equally weighted samples is directly proportional to the variation of the integrand. In order to highlight the relationship between variation and integration error, a perturbation analysis is performed on the integral approximations obtained from two similar point sets. Restricting the analysis to functions in $C^1_{[a,b]}$, consider two point sets $X = \{x_1, \ldots, x_N\} \in [0, 1]$ and $Y = \{y_1, \ldots, y_N\} \in [0, 1]$. Here, $C^1_{[a,b]}$ is the set of all continuous functions with continuous first derivatives over the interval $[a, b]$. Next, suppose these two point sets are similar enough that the difference between any corresponding pair of points is bounded by a constant, that is $|x_n - y_n| < \epsilon$ for $1 \leq n \leq N$. If the bounding constant ϵ is sufficiently small, the difference in the integral approximations sampling X and Y can then be approximated with a first

order Taylor expansion to yield

$$\left|\frac{1}{N}\sum_{n=1}^{N} f(x_i) - \frac{1}{N}\sum_{n=1}^{N} f(y_i)\right| \approx \frac{1}{N}\left|\sum_{n=1}^{N} (x_n - y_n)\frac{\partial f}{\partial x}\bigg|_{x=x_n}\right|.$$

Applying the triangle inequality to bring the absolute value inside the summation, one is able to bound the Taylor expansion

$$\begin{aligned}\frac{1}{N}\left|\sum_{n=1}^{N}(x_n - y_n)\,\frac{\partial f}{\partial x}\bigg|_{x=x_n}\right| &\leq \frac{1}{N}\sum_{n=1}^{N}|x_n - y_n|\left|\frac{\partial f}{\partial x}\bigg|_{x=x_n}\right. \\ &\leq \frac{\epsilon}{N}\sum_{n=1}^{N}\left|\frac{\partial f}{\partial x}\bigg|_{x=x_n}\right. \end{aligned} \tag{3.34}$$

Thus, the difference between integration approximations using similar point sets depends on the magnitude of the integrand's gradients.

It is common in many numerical methods, when the solution is not known, to evaluate the accuracy of the method based on the change in the final result when the numerical solution is perturbed. If the change in the final result is acceptably small, the numerical solution can be taken with greater confidence. If the numerical result is found to vary wildly under small perturbations, then the accuracy of the numerical method is suspect. Based on (3.34), it is clear that the sensitivity of the integration approximation to perturbations depends on the magnitude of the integrand's gradients. As mentioned earlier, a function with larger gradients typically has a larger variation measure. Therefore, approximating the integral of a function with a larger variation measure yields a numerical solution more sensitive to the sample point sets, which indicates a less reliable solution. While the relationship is qualitative at this point, the result (3.34) is consistent with the Koksma-Hlawka inequality. In the next section, an explicit connection between the gradients of a function and its variation is established, upon which, the result (3.34) is found to closely relate to the one dimensional proof of the Koksma-Hlawka inequality in [127] (see Theorem 2.9).

3.3.1 Variation in one dimension

Formally, the variation of a one dimensional function is defined by:

Definition 3.2 *A function f defined on an interval $[a, b]$ is said to be of* bounded variation *if there exists a positive constant C such that*

$$\sum_{n=1}^{N} |f(x_n) - f(x_{n-1})| \leq C \tag{3.35}$$

for every partition

$$a = x_0 < x_1 < \cdots < x_N = b$$

of $[a, b]$ by points of subdivision $x_1, \ldots, x_{N-1}$.

The smallest constant C that satisfies (3.35) is then defined as the variation $V(f)$ of the function f on $[a, b]$.

If the function f is monotonic over the domain $[a, b]$, then the variation is simply the magnitude of the difference between the function evaluated at the endpoints

$$V(f) = |f(b) - f(a)|.$$

Similarly, if the domain $[a, b]$ of the function f can be completely partitioned into subintervals $[a, y_1), [y_1, y_2), \ldots, [y_{k-1}, y_k), [y_k, b]$ such that f is monotonic over each subinterval, then the variation on $[a, b]$ is calculated by the sum of the variation over each subinterval. Hence,

$$V(f) = |f(y_1) - f(a)| + |f(y_2) - f(y_1)| + \cdots + |f(y_k) - f(y_{k-1})| + |f(b) - f(y_k)|. \tag{3.36}$$

As long as the partition points $y_1, y_2, \ldots, y_k$ include every local extrema of the function f in the interval $[a, b]$, the resulting partition satisfies the monotonicity constraint necessary for (3.36). If $f \in C^1_{[a,b]}$, that is to say that f is a continuous function with

a continuous first derivative on $[a, b]$, then the fundamental theorem of calculus can be repeatedly applied to simplify (3.36) and yield

$$\begin{aligned} V(f) &= \int_a^{y_1} \left|\frac{\partial f}{\partial x}\right| dx + \int_{y_1}^{y_2} \left|\frac{\partial f}{\partial x}\right| dx + \cdots + \int_{y_{k-1}}^{y_k} \left|\frac{\partial f}{\partial x}\right| dx + \int_{y_k}^{b} \left|\frac{\partial f}{\partial x}\right| dx \\ &= \int_a^b \left|\frac{\partial f}{\partial x}\right| dx. \end{aligned} \tag{3.37}$$

Thus, the calculation of $V(f)$ on $[a, b]$ requires all the local extrema of f to be found in $[a, b]$ in order to apply (3.36). Alternatively, the variation can be calculated directly from (3.37), when the indicated derivative is well-defined.

Returning to the example of the function $f_1(x; \tau)$ in (3.33), the variation is calculated by first identifying the local extrema of the function. The local extrema of $f_1(x; \tau)$ occur when the cosine argument is an integer multiple of π. More specifically, the maximum value $f_1(x; k) = 2$ is obtained whenever τx^2 equals an even multiple of π, and a minimum value $f_1(x; \tau) = 0$ is obtained whenever τx^2 is equal to an odd multiple of π. The number of extrema located within the interval $(0, 1)$ is equal to $\lceil \tau/\pi \rceil - 1$. Note that each extrema of $f_1(x; \tau)$ in the interval $(0, 1)$ contributes a value of 2 to the sum for the variation in (3.36). After including the contribution of the endpoints, an explicit formula for the variation of $f_1(x; \tau)$ is then given by

$$V\left(f_1(x;\tau)\right) = \begin{cases} 2\left(\lceil \tau/\pi \rceil - 1\right) + f_1(1;\tau) & \text{if } \lceil \tau/\pi \rceil \text{ is even} \\ 2\lceil \tau/\pi \rceil - f_1(1;\tau) & \text{otherwise.} \end{cases} \tag{3.38}$$

Using the result (3.38), the variation is calculated for three representative values of $\tau = 10, 100$, and 1000

$$\begin{aligned} V\left(f_1(x;10)\right) &= 6.161 \\ V\left(f_1(x;100)\right) &= 63.86 \\ V\left(f_1(x;1000)\right) &= 637.6. \end{aligned}$$

In order to better understand the effect of variation on the integration error, the integrals $f_1(x; \tau)$ for $\tau = 10, 100$, and 1000 are approximated using a set E_N of N

points with equal spacing

$$E_N = \left\{ \frac{2n-1}{2N} : 1 \leq n \leq N \right\}. \tag{3.39}$$

The star discrepancy of the set $D_N^*(E_N) = 1/2N$ is the minimum attainable for a one dimensional N-point sequence (see Theorem 2.6 [127]). The relative integration error of the representative functions is given in Figure 3.3(b). For a fixed sequence size, the integration error is found to be directly proportional to the variation which is consistent with the Koksma-Hlawka inequality (3.5). An increase in the variation of $f_1(x;k)$ by an order of magnitude yields an increase in error by roughly an order of magnitude. Since the star discrepancy of E_N converges linearly, the error bound given by the Koksma-Hlawka inequality also converges linearly. However, it is interesting to note that in Figure 3.3(b), the actual numerical results appear to converge quadratically $\mathcal{O}\left(N^{-2}\right)$. The Koksma-Hlawka inequality (3.5) does not consider the periodicity of the integrand which can significantly improve the accuracy of the integration approximation. Thus, it is important to remember that the Koksma-Hlawka inequality only serves as an upper bound on the error, and can be overly pessimistic; especially when the sequence is well-distributed and the integrand is sufficiently periodic.

It is possible for a function to have a finite integral evaluation but an unbounded variation. An obvious example is a monotonic function that rapidly approaches infinity at an endpoint of the integration domain. The function $f_2(x)$ defined on the interval $(0, 1]$ as

$$f_2(x) = \frac{e^{-x}}{\sqrt{x}}, \tag{3.40}$$

is such an example

$$\lim_{x \to 0} f_2(x) \to \infty \quad \text{and} \quad \int_0^1 f_2(x)dx = \sqrt{\pi}\text{erf}(1).$$

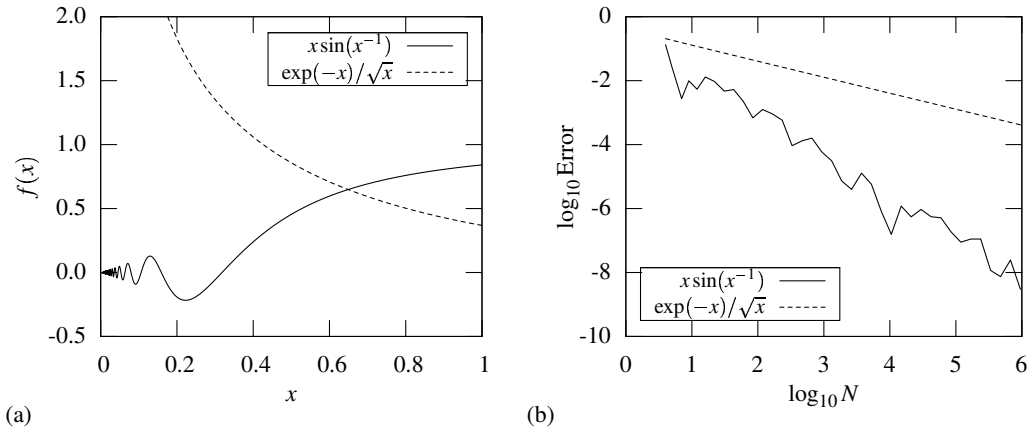

Figure 3.4: Examples of functions with unbounded variation: (a) plot of the functions; and (b) convergence of the integration error.

A plot of the function $f_2(x)$ is given in Figure 3.4(a). While every unbounded function has an unbounded variation, the converse is not necessarily true. Functions that are bounded throughout the integration domain can still possess an unbounded variation.

Consider the function $f_3(x)$ defined on the interval $(0, 1]$ as

$$f_3(x) = x \sin(x^{-1}). \tag{3.41}$$

The function $f_3(x)$ is bounded on the domain,

$$\lim_{x \to 0} f_3(x) = 0,$$

see Figure 3.4(a), and has a definite integral evaluation

$$\int_0^1 f_3(x)dx = \frac{1}{2}\left(\sin(1) + \cos(1) + \mathrm{Si}(1)\right) - \frac{\pi}{4},$$

where $\mathrm{Si}(z) = \int_0^z x^{-1} \sin(x)dx$ is the so-called "sine integral" found in integration tables [1] or calculated using software like Mathematica [189]. On the interval $(0, 1]$ the argument of sine function has a range in $[1, \infty)$ and thus yields an infinite number of local extrema. The presence of an infinite number of extrema alone is not enough to prove that a function has unbounded variation. For example, the function $x^2 \sin(x^{-1})$

has an infinite number of extrema but also a bounded variation on the interval $(0, 1]$. The key is in the magnitude of consecutive extrema. The extrema of the function $\sin(x^{-1})$ are ± 1, and nearly all are located near $x = 0$. As a result, the contribution from these near zero extrema to the total variation is small enough in the case of $x^2 \sin(x^{-1})$ to yield a bounded variation.

To prove a function has an unbounded variation, it is sufficient to show a specific partition of the interval can not be bounded by any constant C in the definition (3.35). Define the following partition $P = \{y_1, y_2, \ldots, y_{2N-1}\}$ for $1 \leq n < 2N$

$$y_n = \begin{cases} \frac{2}{(2n-1)\pi} & \text{if } n \text{ is odd} \\ \frac{2}{n\pi} & \text{if } n \text{ is even} \end{cases}.$$

Note the points of subdivision are decreasing in magnitude $0 < y_{2N-1} < y_{2N-2} \cdots < y_1 < 1$, and are chosen such that when n is odd $\sin(y_n^{-1}) = 1$ and when n is even $\sin(y_n^{-1}) = 0$. The variation measure of f_3 for this specific partition is denoted by $V(f_3; P)$ and has an explicit form

$$V(f_3; P) = \sum_{n=1}^{2N} |f_3(y_n) - f_3(y_{n-1})|,$$

defining $y_0 = 1$ and $y_{2N} = 0$ as the interval endpoints. A simple lower bound is established for $V(f_3; P)$ by considering the consecutive pairs of subdivision points P and omitting the contribution by y_0 and y_1

$$V(f_3; P) > \sum_{n=1}^{N-1} \frac{4}{(4n+1)\pi}. \tag{3.42}$$

However, there is an infinite number of local extrema; and when the partition P is extended by increasing N, the summation in the lower bound of (3.42) is unbounded

$$\lim_{N\to\infty} \sum_{n=1}^{N-1} \frac{4}{(4n+1)\pi} \to \infty.$$

Thus, the variation measure $V(f_3; P)$ is unbounded. If there exists a specific partition of a function that yields an unbounded variation measure, then by definition (3.35) there is no finite constant bounding the variation. Therefore, the function f_3 (3.41) does not have bounded variation on the interval $[0, 1]$.

In an effort to observe the potential consequences of unbounded variation on the integral approximation of a function, the error convergence is given in Figure 3.4(b) for the functions f_2 in (3.40) and f_3 in (3.41). Similar to the examples in Figure 3.3, the integrals approximations are obtained by sampling the functions with the minimum discrepancy point set E_N defined in (3.39). Both functions converge for the number of samples provided, but the unbounded function $f_2(x) = e^{-x}/\sqrt{x}$ converges much slower than the bounded function $f_3(x) = x\sin(x^{-1})$. However, it should be noted that this does not represent a general observation. In fact, the converse is shown to be true for the multi-dimensional test integrals considered in [115, 117, 120, 146]. The convergence of the integral approximation of an unbounded function primarily depends on the strength of the singularity, that is the rate at which the function approaches infinity. The practical lesson to be taken from these functions with unbounded variation is the following. If the Koksma-Hlawka inequality does not bound the integration error because a function is of unbounded variation, then it is difficult to make any assurances regarding the error convergence rate.

3.3.2 Variation in multiple dimensions

Unlike the definition of variation in one dimension, the variation in multiple dimensions does not have a universal definition. This is primarily due to the fact that there are many different ways to partition a multi-dimensional integration domain, and to measure the fluctuations of a function over these partitions. Hence, a specific

definition of multi-dimensional variation must include both the family of partitions and the fluctuation measure over these partitions. The most common definitions of multi-dimensional variation, including those used in this investigation, are named after the mathematician(s) who first proposed or studied them.

For QMC integration, the two important definitions of multi-dimensional variation are the variation in the sense of Vitali and the variation in the sense of Hardy and Krause. The importance of the variation in the sense of Hardy and Krause is evident from its role in the Koksma-Hlawka inequality. The variation in the sense of Vitali is important as it is used to calculate the variation in the sense of Hardy and Krause. For both types of multi-dimensional variation, the only family of partitions considered are those consisting solely of subintervals. Here, a subinterval in s dimensions is simply another way to refer to a hyper-rectangle of the form $\prod_{i=1}^{s}[a_i, b_i)$, where $0 \leq a_i < b_i \leq 1$ for $1 \leq i \leq s$. The complete definitions for the multi-dimensional variations of Vitali, and Hardy and Krause are provided in the paragraphs that follow. While multi-dimensional variation is important to the study of QMC methods, it is somewhat difficult to find introductory analysis texts that cover the concept of variation beyond one dimension. Owen [136], in his Stanford Department of Statistics technical report, reviews many important properties of multi-dimensional variation; in particular, as it applies to QMC integration.

The definitions provided in this investigation for multi-dimensional variation follow closely those of Niederreiter in [127]. For a function $f(\mathbf{x})$, with $\mathbf{x} \in \overline{I}^s$, the *variation of f in the sense of Vitali* is defined by

$$V^{(s)}(f) = \sup_{P \in \mathcal{P}} \sum_{J \in P} |\Delta(f; J)| , \tag{3.43}$$

where $\mathcal{P}$ is the family of all partitions over $\overline{I}^s$ into subintervals, J is a specific subin-

terval from a given partition $P \in \mathcal{P}$. Here, $\Delta(f; J)$ denotes the alternating sum of the values of f evaluated at the vertices of the subinterval; that is to say that function values at adjacent vertices have opposite signs. For example, given a function $f(x_1, x_2)$ over two dimensions, and the subinterval $j = [a_1, b_1) \times [a_2, b_2)$ then

$$\Delta(f; J) = f(a_1, a_2) - f(a_1, b_2) + f(b_1, b_2) - f(b_1, a_2).$$

It is important to note because of the absolute value in (3.43), the actual sign choice of a particular vertex is not important, only that adjacent vertices have opposite signs.

In an analogous manner to the one dimensional result in (3.37), a more explicit form for the variation in the sense of Vitali,

$$V^{(s)}(f) = \int_0^1 \cdots \int_0^1 \left| \frac{\partial^s f}{\partial u_1 \cdots \partial u_s} \right| du_1 \cdots du_s, \tag{3.44}$$

is applicable whenever the indicated partial derivative is continuous on the integration domain. Regardless of the means by which the variation is calculated, if $V^{(s)}(f)$ is found to be finite, then the function f is said to be of bounded variation in the sense of Vitali.

In order to construct the definition of variation in the sense of Hardy and Krause, some additional notation is necessary. For $1 \leq k \leq s$ and $1 \leq i_1 < i_2 < \cdots < i_k \leq s$, let $V^{(k)}(f; i_1, \ldots, i_k)$ denote the variation in the sense of Vitali of the restriction of f to the k dimensional face F of the unit hypercube given by

$$F = \left\{ (u_1, \ldots, u_s) \in \overline{I}^s \ : \ u_j = 1 \text{ for } j \neq i_1, \ldots, i_k \right\}.$$

Then the *variation of f in the sense of Hardy and Krause* is defined by

$$V_{HK}(f) = \sum_{k=1}^{s} \sum_{1 \leq i_1 \leq i_2 \leq \cdots \leq i_k \leq s} V^{(k)}(f; i_1, \ldots, i_k). \tag{3.45}$$

Keeping with prior convention, if $V_{HK}(f)$ is found to be finite, then the function f is said to be of bounded variation in the sense of Hardy and Krause.

Morokoff in [117] questions the effect variation has on the integration error based on the results approximating several multi-dimensional test integrals. One particular example from [117], that casts doubt on the role of variation, involves two nearly identical functions over $\mathbf{x} \in \overline{I}^s$:

$$\begin{aligned} g(\mathbf{x}) &= \prod_{i=1}^{s} x_i, \text{ and} \\ h(\mathbf{x}) &= \prod_{i=1}^{s} (1 - x_i). \end{aligned} \tag{3.46}$$

With a simple variable transformation, it is clear that the two functions $g(\mathbf{x})$ and $h(\mathbf{x})$ have the same integral value over $\overline{I}^s$. Furthermore, given the problem symmetry, it should be expected that both integral approximations should have the same expected accuracy. The Koksma-Hlawka inequality, however, does not imply that the accuracy will be the same because the variation in the sense of Hardy and Krause differs greatly in the multi-dimensional case. Specifically, $V_{HK}(g) = 2^s - 1$ and $V_{HK}(h) = 1$. The variation in the sense of Hardy and Krause is constant with dimension for the function h because the function has a constant value of zero on every face of the unit hypercube whenever $x_i = 1$ for at least one dimension $1 \leq i \leq s$. As a result, all the terms in the summation in (3.45) except $V^{(s)}(f)$ are zero. Again, it is important to remember that the Koksma-Hlawka inequality only serves as an upper bound to the integration approximation error.

Using a similar two dimensional example as in (3.46), the integral approximations to the functions

$$\begin{aligned} f_4(x, y) &= (x - 1)^2 (y - 1)^2, \text{ and} \\ f_5(x, y) &= x^2 y^2, \end{aligned} \tag{3.47}$$

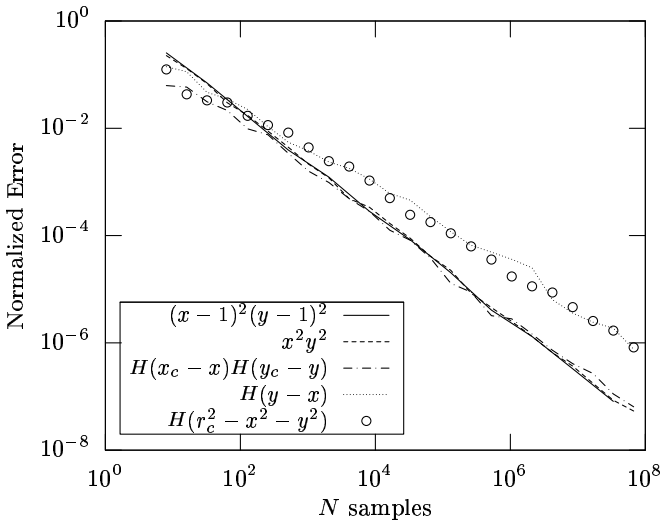

Figure 3.5: Integration error of several test functions using a two dimensional Halton sequence.

are found using the two dimensional low-discrepancy Halton sequence (see Appendix C). Note that $V_{HK}(f_4) = 1$ and $V_{HK}(f_5) = 3$. In Figure 3.5, the error convergence rate for both functions is nearly linear as expected from a low-discrepancy sequence; but more importantly, the actual integration error of both functions is almost identical. Hence, the expected result is obtained, even though the Koksma-Hlawka inequality suggests that it would be possible for the function f_5 to have up to 3 times greater integration error than f_4. In order to best assess the performance with the low-discrepancy Halton sequence in Figure 3.5, the error given represents the average of 16 calculations of the integral approximation using consecutive subsequences for each function.

The effect of variation on the expected integration error becomes more important when the integrands contain discontinuities. Discontinuous integrands are very important to many types of Monte Carlo simulation because they occur whenever a yes/no decision is made in the simulation. Unfortunately, when the discontinuities of a function are not aligned with the partition boundaries in (3.43), the resulting

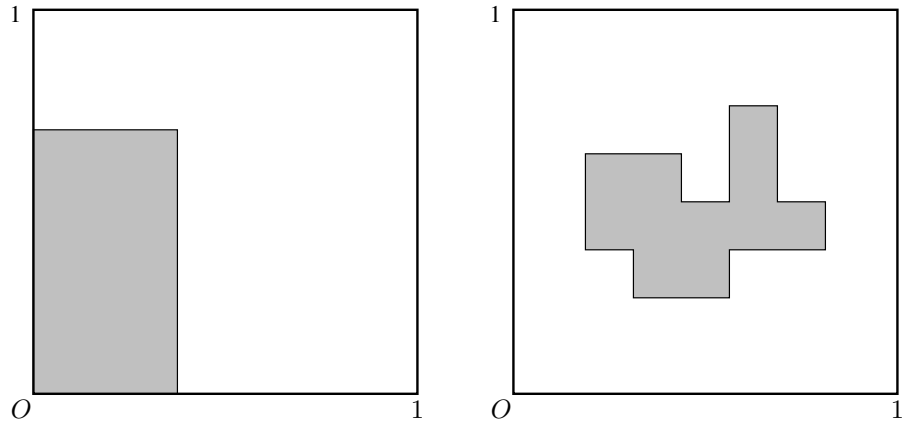

Figure 3.6: Bounded variation in the sense of Vitali of two simple indicator function on $[0,1)^2$: (left) $V^{(2)}(f) = 1$; and (right) $V^{(2)}(f) = 14$.

multi-dimensional variation in the sense of Vitali is not bounded. Moreover, if a function is not of bounded variation in the sense of Vitali (and by extension, Hardy and Krause), then the Koksma-Hlawka inequality does not bound the integration error of the function. Without the Koksma-Hlawka inequality, little can be established regarding the expected error convergence rate. For the variation definitions of Vitali and Hardy and Krause, all the partition boundaries are aligned with principal axes. Thus, if a function possesses a discontinuity that is not aligned with the principle axes, it will not be of bounded variation in the sense of Hardy and Krause.

In Figure 3.6, two examples of discontinuous indicator functions are given along with their associated variation in the sense of Vitali. Here, the shaded region represents a function value of one and the non-shaded region represents a function value of zero. Since, their discontinuities are aligned with the primary axes (x and y), their variation in the sense of Vitali is bounded. Furthermore, since the function in both cases is constant along the boundaries $x = 1$ and $y = 1$, the variation in the sense of Vitali equals the variation in the sense of Hardy and Krause. The function

$$f_6(x, y) = H(x_c - x)H(y_c - y),$$

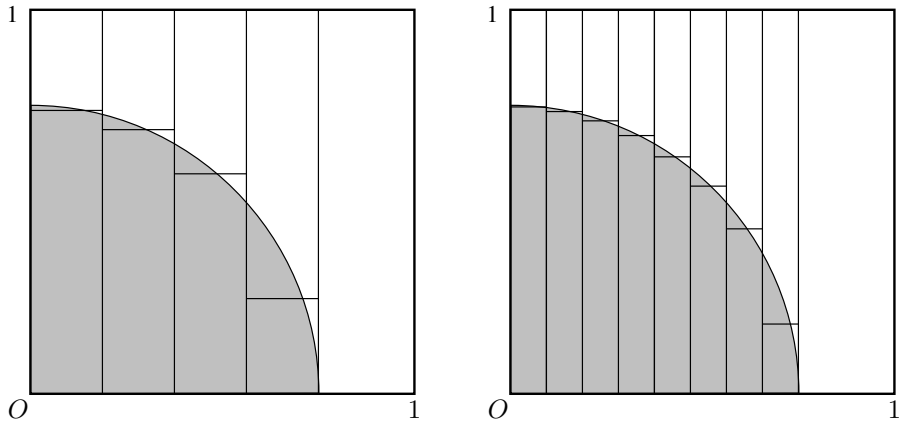

Figure 3.7: Variation of an indicator function f on $[0,1)^2$ using different domain partitions: (left) $V(f;P_4)=\frac{17}{2}$; and (right) $V(f;P_8)=\frac{33}{2}$.

is representative of a discontinuous multi-dimensional function with bounded variation as illustrated in Figure 3.6(a). Here $H(x)$ represents the Heaviside step function defined by

$$H(x)=\begin{cases}1 & \text{if } x>0\\ \frac{1}{2} & \text{if } x=0\\ 0 & \text{if } x<0.\end{cases}.$$

For $x_c=y_c=2^{-1/2}$, the convergence of the integration error for f_6 is plotted in Figure 3.5. Not only is the error convergence rate of f_6 nearly linear, it is almost the same magnitude as the two continuous functions f_4 and f_5 from (3.47).

In order to understand why a function with discontinuities not aligned with the principle axes is not of bounded variation in the sense of Vitali, consider the circular indicator function

$$f_7(x,y)=H(r_c^2-x^2-y^2), \tag{3.48}$$

given in Figure 3.7. Now, if f_7 is of bounded variation, then there must exist some constant that bounds the summation in (3.43) for every possible partition. However, this is not the case, because one can construct a simple family of partitions such

that the summation term in (3.43) is shown to grow without bound. First subdivide the circle radius along the x axis into M subintervals. Then divide each of these subintervals into two more subintervals such that the boundary between these subintervals intersects the discontinuity boundary, see Figure 3.7 for examples of these partitions.

To state this more concretely, let P_M represent the partition of the integration domain explicitly defined by

$$\begin{aligned} P_M = \Big\{ & \left[0, \tfrac{1}{M} r_c\right) \times [0, y_1), \left[\tfrac{1}{M} r_c, \tfrac{2}{M} r_c\right) \times [0, y_2), \ldots, \left[\tfrac{M-1}{M} r_c, r_c\right) \times [0, y_M), \\ & \left[0, \tfrac{1}{M} r_c\right) \times [y_1, 1), \left[\tfrac{1}{M} r_c, \tfrac{2}{N} r_c\right) \times [y_2, 1), \ldots, \left[\tfrac{M-1}{M} r_c, r_c\right) \times [y_M, 1), \\ & [r_c, 1) \times [0, 1) \Big\}. \end{aligned}$$

Here y_n is the point where the subinterval boundaries intersect the discontinuity boundary at

$$y_n = \frac{1}{2}\left(\sqrt{r_c^2 - \left(\tfrac{n-1}{N}\right)^2} + \sqrt{r_c^2 - \left(\tfrac{n}{N}\right)^2}\right), \text{ for } 1 \le n \le M.$$

Let $V^{(s)}(f; P_M)$ represent the evaluation of the summation in (3.43) for the specific partition P_M. Hence,

$$V^{(s)}(f; P_M) = \sum_{J \in P_M} |\Delta(f; J)|,$$

where J is a subinterval of P_M. For the circular indicator function f_7 in (3.48),

$$V^{(s)}(f_7; P_M) = M + \tfrac{1}{2},$$

which grows without bound as M increases. Note that $V^{(s)}(f; P_M)$ is a lower bound for the variation in the sense of Vitali. Therefore, the function f_7 is not of bounded variation in the sense of Vitali; and thus by extension, it is also not of bounded variation in the sense of Hardy and Krause.

Now consider the function f_7 in (3.48), and f_8 defined by

$$f_8(x, y) = H(y - x), \tag{3.49}$$

which both possess a discontinuity not aligned with the principle axes. The integration error using the low-discrepancy Halton sequence to sample the discontinuous functions f_7 and f_8 are plotted in Figure 3.5. Note for the circular indicator function f_7 that $r_c = 2/\pi$. In both cases, the error convergence rate is approximately $\mathcal{O}\left(N^{-3/4}\right)$, which yields a greater error for the same number of samples ($N > 128$) than the functions with bounded variation. Thus, it appears that when a function is not of bounded variation that the error convergence rate is negatively impacted.

As illustrated by the examples of Morokoff [117], the connection between a function's variation and its associated integration error, with respect to the Koksma-Hlawka inequality, is questionable. When a function is not of bounded variation in the sense of Hardy and Krause, the Koksma-Hlawka inequality does not provide any information about the error convergence. This does not mean the integral approximation will not converge in practice, only that the unbounded variation prevents an upper error bound from being established. Despite the lack of a direct link between a function's variation and the magnitude of the error in its integral approximation, unbounded variation is typically a strong indicator that the error convergence rate will be less than the star discrepancy. It should be stressed that when a function possesses unbounded variation due to discontinuities not aligned with the principle axes, the resulting error convergence rate of the integral approximation is negatively impacted. This is supported by the preceding examples in (3.48) and (3.49), and the discontinuous test integrals in [115, 117, 120, 146]. Moskowitz, in [120], investigates smoothing techniques to eliminate the discontinuities directly from an integrand and

also the discontinuities that are introduced when the acceptance-rejection sampling technique is used. In general, such discontinuous integrands should be modified or avoided when constructing a QMC method in order to achieve the best possible error convergence rate.

3.4 Low discrepancy sequences versus optimal integration lattices

Now that the concepts of discrepancy and variation have been discussed in the context of the Koksma-Hlawka inequality, it is time to focus on finding point sets P that are low-discrepancy, that is $D^*(P) \approx \mathcal{O}\left(N^{-1}(\log N)^{s-1}\right)$. In the sample star discrepancy calculations presented in Figure 3.2, two types of low-discrepancy point sets are briefly introduced: the Halton sequence, and the Korobov lattice. The Halton sequence is a specific example of a low-discrepancy sequence, and the Korobov lattice is a specific example of an optimal integration lattice. The low-discrepancy sequence and the optimal integration lattice constitute two fundamental classes of low-discrepancy point set design. Typically, for the same number of points, it is possible to find an optimal integration lattice with a lower discrepancy than a sequence. However, in practice, the low-discrepancy sequences are preferred for quasi-Monte Carlo integration when applied to particle-type problems. The two design classes are compared for the remainder of this section, and the practical advantages of low-discrepancy sequences, as they apply to particle problems in this investigation, are highlighted.

A sequence, in general, is a set of points $P = (x_m, x_{m+1}, \ldots)$ such that each point x_n is defined by its position n for $n \geq m$ (where m is typically zero or one) [149]. A low-discrepancy sequence is a sequence of points constructed in such a manner that

the star-discrepancy of the first N points of the sequence is as low as practicable for all $N \geq 0$. It should be noted that this does not mean that the star discrepancy of a low-discrepancy sequence will monotonically decrease with N. At some values of N during the construction of a low-discrepancy sequence, it is necessary for the star discrepancy to increase in order to ensure a lower discrepancy for larger values of N. These increases are small enough to maintain an overall asymptotic convergence of the star discrepancy of the first N points of a low-discrepancy sequence to be $D^*(P) \approx \mathcal{O}\left(N^{-1}(\log N)^{s-1}\right)$. Hence, the general trend for a low-discrepancy sequence is to have its star discrepancy to converge nearly linearly to zero.

As an example, the convergence of the star discrepancy of the low-discrepancy van der Corput sequence in base 2 is given in Figure 3.8. Note that the star discrepancy of the van der Corput sequence does not monotonically decrease. However, the star discrepancy does achieve a near linear convergence rate for any sequence length, which is superior to a random sequence. Also note in Figure 3.8, that at regularly spaced intervals on the logarithmic scale, the star discrepancy of the van der Corput sequence has local minima nearly an order of magnitude less than similar sequence lengths. These minima correspond to sequence lengths that are a power of 2, and are a consequence of the construction of the van der Corput sequence in base 2. The construction of low-discrepancy sequences is discussed in much greater detail in Chapter IV.

In contrast to the low-discrepancy sequence, given an N point optimal integration lattice L, any strict subset of L is not required to be an optimal integration lattice. Consider the following one dimensional case, the point set $P = (x_1, \ldots, x_N)$ achieves the minimum possible star discrepancy of $D_N^*(P) = 1/2N$ [127] when the points of

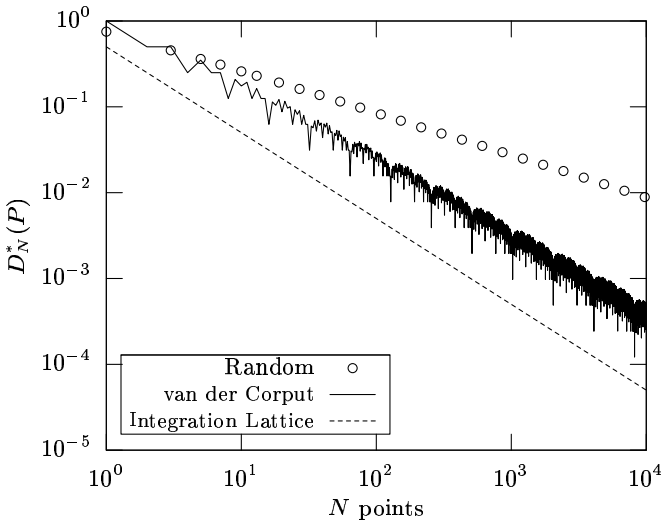

Figure 3.8: Illustration of the star discrepancy $D_N^*(P)$ convergence for different one-dimenisional point sets P.

P are given by

$$x_n = \frac{2n-1}{2N} \quad \text{for } 1 \leq n \leq N, \tag{3.50}$$

which represents an optimal integration lattice. Note that the set of all optimal one dimensional lattices is equivalent to the family of sets E_N, for $N = 1, 2, \ldots$, defined in (3.39), which accounts for the super-linear error convergence observed in Figures 3.3 and 3.4. Since an optimal integration lattice is not required to reuse the points found for smaller lattices, there are fewer restrictions on the selection of points in P compared to a low-discrepancy sequence. As a result one would expect that a lower discrepancy could be achieved with an N point optimal integration lattice compared to a low-discrepancy sequence of the same size. For example, in Figure 3.8, the optimal one dimensional integration lattice in (3.50) is indeed found to possess a lower discrepancy than the low-discrepancy van der Corput sequence for any size point set. It should be noted that one can not select just any points for an optimal integration lattice, there are some conditions from group theory that apply to the definition of an integration lattice. However, these conditions are not nearly as

restrictive as those placed on the construction of a low-discrepancy sequence. More information on the formal definition of an integration lattice can be found in great detail in [127, 160].

While having fewer restriction than a low-discrepancy sequence, an optimal lattice does has more restrictions than a general point set. However, it would be impractical to try and find a general low-discrepancy point set in most cases. The star discrepancy $D_N^*(P)$ of an s-dimensional point set $P = (\mathbf{x_1}, \ldots, \mathbf{x_N})$ is a continuous function with respect to each coordinate of each point in the set P [127], but its derivative is not well-defined everywhere in the domain. While a function minimization routine based on the local gradient, such as the method of steepest descent [147], is not applicable, it would be possible to use a general minimum search. However, for the number of points typically needed for accurate integration, the large dimensions of the minimization problem would limit the possible search techniques to the simple Monte Carlo search. In this case, one would essentially generate random point sets which are known to have a poor expected star discrepancy $D_N^*(P) = \mathcal{O}\left(N^{-1/2}(\log\log N)^{1/2}\right)$; and hope to find one by luck that yields an extremely low discrepancy. Furthermore, even this simple Monte Carlo search method requires the discrepancy to be evaluated for each candidate point set; and from Section 3.2, the cost of calculating the exact value of the discrepancy for even a single point set is extremely prohibitive.

As mentioned earlier, one of the advantages of the optimal integration lattice is that often it is possible, at least in theory, to find a lattice with a lower star-discrepancy than a low-discrepancy sequence with the same number of points. In addition to this potential advantage, an optimal integration lattice is also able to exploit the regularity of a periodic integrand to achieve a super-linear error convergence rate. Here, the regularity of a function $f(\mathbf{x})$, with $\mathbf{x} \in \overline{I}^s$, refers to the rate at

which its Fourier coefficients, defined by

$$\hat{f}(\mathbf{h}) = \int_{\bar{I}^s} f(\mathbf{x}) e^{2\pi i \mathbf{x}\cdot\mathbf{h}} d\mathbf{x} \quad \text{for all } \mathbf{h} \in \mathbb{Z}^s,$$

converge toward zero. Specifically, let $\alpha > 1$ and $C > 0$ represent real numbers such that

$$\left|\hat{f}(\mathbf{h})\right| \leq C \left(\prod_{i=1}^{s} \max(1, h_i)\right)^{-\alpha}, \tag{3.51}$$

then α is said to describe the regularity condition on the function f. The larger the value of α for an integrand, the faster the approximation using an optimal integration lattice will converge.

It is possible to establish for optimal integration lattices a better convergence rate than for low-discrepancy sequences because the lattice error lends itself well to Fourier analysis, which provides a tighter bound than the Koksma-Hlawka inequality. However, the quadratic (and higher) convergence possible with the optimal integration lattices only occurs for a special class of functions when $\alpha \geq 2$. It should be noted that this condition is much more restrictive than being of bounded variation in the sense of Hardy and Krause. As such, the Koksma-Hlawka inequality applies to many more types of functions than just those able to achieve super-linear convergence with an optimal integration lattice.

The general advantage of using a low-discrepancy sequence is that the actual building of the sequence is a constructive process. That is to say a low-discrepancy sequence can be made any length and dimension by repeating a few relatively simple steps. Furthermore, if one has already calculated the first N points of a low-discrepancy sequence in s dimension, these are retained if one wishes to increase the sequence length or the problem dimension. In contrast, the optimal integration lattices while proven to exist, can not be built in a constructive manner. For

dimensions greater than 2, one must generally use an exhaustive search to find an optimal integration lattice with N points.[6] Without any means to guide the search for an optimum lattice, one is forced to check every possible lattice in an exhaustive search, which is a non-constructive process. Thus, if more lattice points are needed to improve the accuracy, or if the problem dimension increases, the current optimal lattice found must be discarded and the entire exhaustive search repeated.

In spite of its potential for a lower discrepancy and super-linear convergence, the low-discrepancy sequence is actually preferred over the optimal integration lattice for particle-type QMC integration. There are two main reasons why the low-discrepancy sequence is preferred in this case. The first reason is that for many particle simulations it is not known *a priori* how many samples are necessary to achieve an acceptable error. If one were to generate a fixed number of samples in a QMC particle simulation, but found the accuracy was unsatisfactory, one could simply forge ahead and continue the low-discrepancy sequence by adding new samples as necessary with minimal computational effort. However, the same could not occur for the optimal integration lattice which must be discarded and a new larger lattice would have to be found in its place.

The second reason is that the size of the point sets needed for a QMC particle simulation are generally much larger than the size of the optimal integration lattices that can be found in practice. For example, the particle simulations presented in Chapter VI use low-discrepancy discrepancy point sets containing 2^{26} points in 300 dimensions. Generating these points from any of the low-discrepancy sequences considered in Section 4.3 can be achieved in a matter of a couple minutes on a single 3GHz processor. Unfortunately, finding an optimal integration lattice of the same

[6]This is an active area of research within the QMC community with Sloan [160], and Dick and Kuo [42], among others investigating more constructive integration lattice designs.

size, ignoring the memory limitations, would require several lifetimes using a modern desktop computer. Simple search algorithms for lattices with the special form of Korobov [82, 83] require a minimum of $\frac{1}{2}N^2s$ operations, where N is the number of points in an s-diemnsional lattice [106].[7] More general search algorithms are often combinatorially in nature and thus the computational cost grows exponentially with the lattice dimension (see the algorithms in [17, 70, 104, 106]).

Furthermore, for most particle simulations like DSMC, the whole purpose of adopting the particle formulation is to avoid solving the integrand directly. Without the explicit form of the integrand available, it is difficult to properly periodize the integrand using the techniques in [160] to ensure that the regularity condition (3.51) is satisfied. If the integrand does not meet the necessary regularity condition, then the super-linear convergence of the optimal integration lattice is not assured. Given the practical advantages of low-discrepancy sequences over optimal integration lattices for QMC particle simulations, only low-discrepancy sequences are considered for this investigation. Their construction and performance in QMC integration is detailed in the following chapter.

[7] Specifically, at least $6 \cdot 10^{17}$ opeartions are required to find the optimum Korobov integration lattice that is the same size as the largest low-discrepancy sequence used in this investigation ($N = 2^{26}$ and $s = 300$).

CHAPTER IV

LOW-DISCREPANCY SEQUENCES

The main purpose of this chapter is to review the algorithms used in this investigation to generate the pseudo-random and low-discrepancy sequences needed for the Monte Carlo and quasi-Monte Carlo (QMC) particle simulations. A low-discrepancy sequence is a deterministic sequence of points ($P = \mathbf{x}_0, \mathbf{x}_2, \ldots$), with a star discrepancy that asymptotically approaches zero as rapidly as possible. From the lower bound of Roth [150], the fastest theoretical convergence of the star discrepancy of any sequence is nearly linear. That is,

$$D_N^*(P) > C_s N^{-1} (\log N)^{(s-1)/2}, \tag{4.1}$$

for some constant C_s that depends on the sequence dimension $s \geq 2$. No sequence is actually known to achieve the lower bound of Roth (4.1); however, there are many sequence constructions known to achieve a slightly slower convergence rate that is still nearly linear. If the star discrepancy of a sequence converges to zero at least as fast as $D_N^*(P) = \mathcal{O}\left(N^{-1+\epsilon}\right)$ for all $\epsilon > 0$, then the sequence is referred to as a low-discrepancy sequence. There are four types[1] of s-dimensional low-discrepancy sequences tested in this investigation for the QMC method: (i) the

[1]For reference, the van der Corput sequence and the Sobol' sequence are also included in Appendices A and D, respectively.

Weyl-Richtmyer sequence ($D_N^* = \mathcal{O}\left(N^{-1+\epsilon}\right)$ for all $\epsilon > 0$ – see Appendix B); (ii) the Halton sequence ($D_N^* = \mathcal{O}\left(N^{-1}(\log N)^s\right)$ – see Appendix C); (iii) the Faure sequence ($D_N^* = \mathcal{O}\left(N^{-1}(\log N)^s\right)$ – see Appendix E); and (iv) the Niederreiter sequence in base 2 ($D_N^* = \mathcal{O}\left(N^{-1}(\log N)^s\right)$ – see Appendix F).

As noted in Chapter III, the Koksma-Hlawka inequality (3.5) establishes an error bound on the integral approximation obtained for a general set of points P used to sample the integrand, which is a function of the star-discrepancy of P. In particular, for a function f of bounded variation in the sense of Hardy and Krause, the error in the integral approximation of f converges to zero at the same rate as the star discrepancy. When a low-discrepancy sequence is used to generate the sample points for the integrand, the integral approximation therefore has a near-linear convergence rate as well. This type of numerical approximation is then referred to as quasi-Monte Carlo integration. The theoretical error convergence rate of the QMC method is superior to the $\mathcal{O}\left(N^{-1/2}\right)$ convergence rate associated with the Monte Carlo method; and this is the motivating factor for developing a particle simulation based on the method. A better integral approximation is obtained from the QMC method because the points used to sample the integrand are more evenly-distributed throughout the domain than the random sequence used in the Monte Carlo method. To illustrate this difference in the point distribution of the two sequences, the first 256 points of the low-discrepancy Halton sequence and the pseudo-random sequence are plotted in Figure 4.1. As noted by Press and Teukolsky in [146], each new element of a low-discrepancy sequence is added to the integration domain at a location that "maximally avoids" all the previous sequence elements, thereby producing a distribution of points that is much more uniform than random. In contrast, each element of the pseudo-random sequence is generated independently of the previous

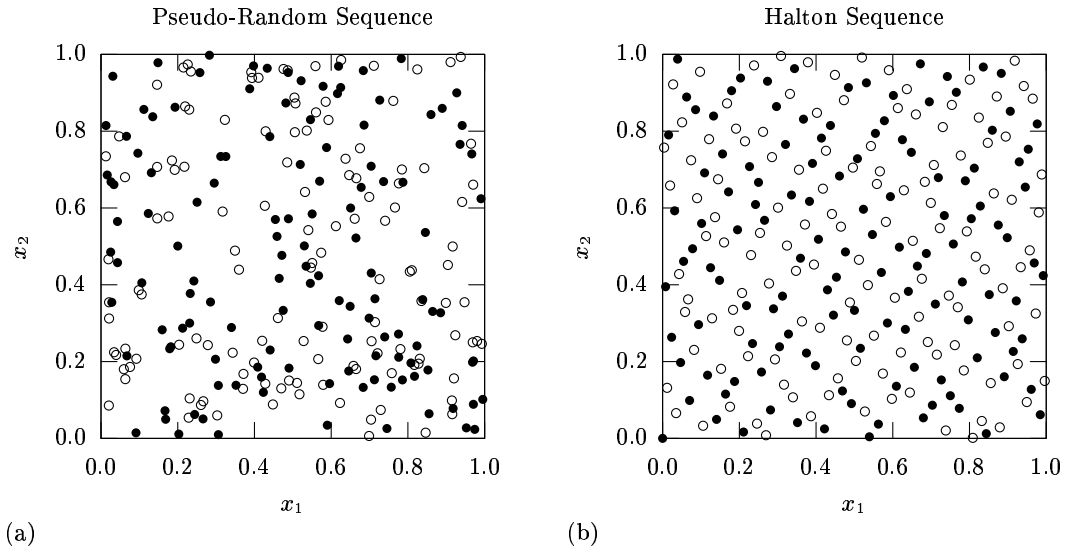

Figure 4.1: The first 256 points of a two dimensional sequence: (a) the pseudo-random sequence; and (b) the Halton sequence in prime bases $p_1 = 2$ and $p_2 = 3$. Note a filled circle denotes one of the first 128 elements of the sequence and an open circle denotes one of the second 128 elements.

elements, resulting in regions of the domain that are both much more sparsely and densely populated than average.

To briefly summarize, the Weyl-Richtmyer sequence is perhaps the simplest low-discrepancy sequence to actually construct; however, the sequence is the least popular in the QMC literature and there is no widely accepted standard for its implementation. As a consequence, a special construction of the Weyl-Richtmyer low-discrepancy sequence, termed the BCF-3 sequence, is proposed in Section 4.1 based on heuristic arguments that suggest it is particularly well-suited for the QMC particle simulations developed here. The actual process for constructing the BCF-3 sequence is outlined, along with examples of the constructive elements of the sequence, in Section 4.2. Finally, in Section 4.3, the algorithmic implementation of the pseudo-random sequence and the four low-discrepancy sequences is discussed and a comparison of the computation time for generating the sequences is presented.

4.1 A Special Construction of the Weyl-Richtmyer Sequence

The Weyl-Richtmyer sequence is the oldest low-discrepancy sequence found in literature; however, it rarely appears in modern applications of the QMC method. In [186], Weyl first defines the sequence and proves that it satisfies certain desirable uniformity conditions on the distribution of its points (see Theorems B.1 and B.2). These uniformity conditions imply that the Weyl-Richtmyer sequence $\mathbf{x}_1, \mathbf{x}_2, \ldots \in \overline{I}^s$ generates a mathematically consistent approximation for the integral of a well-behaved function $f(\mathbf{u})$ with $\mathbf{u} \in \overline{I}^s$, when the sequence is used to sample $f(\mathbf{u})$. That is,

$$\int_{\overline{I}^s} f(\mathbf{u})du = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(\mathbf{x}_n), \tag{4.2}$$

where f is Riemann integrable. This important connection between the distribution of a sequence and its corresponding integral approximation ultimately led to the development of the more formal Koksma-Hlawka inequality [66, 81], which, as noted in Chapter III, serves as the cornerstone of the QMC method. Richtmyer, in [148], proves that the sequence of [186] is able to achieve near-linear error convergence under certain specified conditions[2] when used for the integral approximation in (4.2). In addition to the theoretical convergence results, Richtmyer [148] implements the sequence of [186] in the first numerical QMC simulation, and first coins the term "quasi-Monte Carlo method." It should be noted that there is no "official" name for the sequence originally defined in [186] that appears on a consistent basis throughout the literature.[3] For convenience, the sequence is referred to as the "Weyl-Richtmyer

[2]The proof of Richtmyer in [148] assumes the function to be integrated has an absolutely convergent Fourier series. Additional results are also developed for the special case when the Weyl-Richtmyer sequence is constructed from a set of algebraic irrational numbers that are linearly independent over the rationals.

[3]Examples can be found in literature referring to the same sequence as the "Weyl sequence" [41] and the "Richtmyer sequence" [68], while still others leave the sequence nameless [127].

sequence" throughout this investigation, in recognition of the contributions of both Weyl [186] and Richtmyer [148].

The actual construction for the elements of the Weyl-Richtmyer sequence is remarkably simple compared to the other types of low-discrepancy sequences encountered in this investigation (see Appendices B-F for more details). An s-dimensional Weyl-Richtmyer sequence is defined by an ordered set $\mathbf{z} = (z_1, \ldots, z_s)$ of s irrational numbers that are linearly independent over the rationals.[4] The n^{th} element of the s-dimensional Weyl-Richtmyer sequence $S(\mathbf{z}) = \mathbf{x}_0, \mathbf{x}_1, \ldots \in \overline{I}^s$ is then defined as

$$\mathbf{x}_n = ([nz_1], [nz_2], \ldots, [nz_s]) \in \overline{I}^s, \tag{4.3}$$

where $[\cdot]$ denotes the fractional part of the argument; expressed alternatively, $[x] = x - \lfloor x \rfloor$ where $\lfloor \cdot \rfloor$ is the standard floor function. Once the set $\mathbf{z}$ is selected, it is easy to design an algorithm to generate the Weyl-Richtmyer sequence $S(\mathbf{z})$ using (4.3), which is discussed further in Section 4.3. Based on a result of Niederreiter [123], the extreme discrepancy of the Weyl-Richtmyer sequence $D_N(S(\mathbf{z})) = \mathcal{O}\left(N^{-1+\epsilon}\right)$ for all $\epsilon > 0$, when the set $\mathbf{z}$ is constructed from algebraic irrational numbers that satisfy the linear independence criterion. The Weyl-Richtmyer sequences considered in this investigation are all constructed from quadratic irrational numbers (*i.e.* real numbers that contain a square root of a square free integer). Consequently, due to the result in [123], they are considered low-discrepancy sequences because the extreme discrepancy has near-linear convergence to zero as the sequence length N tends to infinity.[5]

[4] A set $(z_1, \ldots, z_s)$ is defined as *linearly independent over the rationals* if there is no non-trivial solution to the equation $a_1 z_1 + \cdots + a_s z_s = 0$ when $a_1, \ldots, a_s \in \mathbb{Q}$.

[5] While the convergence of the discrepancy of the Weyl-Richtmyer sequence is still near-linear, the rate of convergence is $\mathcal{O}\left(N^{-1+\epsilon}\right)$ for all $\epsilon > 0$ is slightly slower in an asymptotic sense than the $\mathcal{O}\left(N^{-1}(\log N)^s\right)$ convergence found for the s-dimensional Halton, Faure, Sobol', and Niederreiter sequences.

In order to discuss the discrepancy in the special case of a one dimensional Weyl-Richtmyer sequence, it is necessary to introduce the concepts of a continued fraction and a simple continued fraction.

Definition 4.1 *The* continued fraction *is a representation of a real number z by a sequence, possibly infinite, of nested fractions; specifically,*

$$z = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\cdots \cfrac{}{a_{n-1} + \cfrac{1}{a_n}}}}}, \tag{4.4}$$

where the coefficients $a_0, \ldots, a_n$ are real-valued, and the coefficients $a_1, \ldots, a_n$ are strictly positive.

It is rather cumbersome to express a continued fraction explicitly in its full form; therefore, in this investigation, the following compact notation is adopted for the continued fraction in (4.4)

$$z = \langle a_0, a_1, a_2, \ldots, a_n \rangle.$$

Definition 4.2 *The* simple continued fraction *is a continued fraction representation of a real number $z = \langle a_0; a_1, a_2, \ldots, a_n \rangle$ with all integer coefficients.*

As is common in literature, the term *continued fraction* is used throughout this investigation to refer to both types in Definition 4.1 and Definition 4.2 for ease of reading. In fact, every continued fraction that is found here can be assumed to be a simple continued fraction, except where explicitly stated otherwise. The basic

properties of continued fractions are covered in [78, 131], and a more thorough review of the mathematical theory is available in [75].

Suppose there exists an irrational number z and a constant integer k such that the continued fraction of $z = \langle a_0, a_1, \ldots \rangle$ has coefficients that satisfy $a_i \leq k$ for all i. In this special case, there exists a known upper bound on the extreme discrepancy of the one dimensional Weyl-Richtmyer sequence $S(z)$ generated from z. More specifically (see Corollary 3.5 [127]),

$$D_N\left(S(z)\right) < G(k)\frac{\log(N+1)}{N} \quad \text{for all } N \geq 1, \tag{4.5}$$

where $G(k) = \frac{2}{\log 2}$ for $k = 1, 2, 3$, and $G(k) = \frac{k+1}{\log(k+1)}$ for $k \geq 4$. The one dimensional Weyl-Richtmyer sequence $S(z)$ therefore achieves a smaller upper bound on its extreme discrepancy when the coefficients of the continued fraction of z are smaller. From the theory of continued fractions [75, 131], the irrational number z is said to be poorly approximated by the rationals when the coefficients are relatively small in the continued fraction representation of z.[6] The Koksma-Hlawka inequality (3.5) therefore suggests that a better one dimensional integral approximation is expected when the integrand is sampled by a Weyl-Richtmyer sequence $S(z)$, where z is poorly approximated by the rationals.

In general, the specific constructive elements used to generate a low-discrepancy sequence affect the implied constant in the asymptotic bound on the discrepancy of the sequence. Unfortunately, there is no currently known method to determine the constant $C(\mathbf{z})$ in the discrepancy bound $D_N(S(\mathbf{z})) < C(\mathbf{z})N^{-1+\epsilon}$ (for all $\epsilon > 0$) for the multi-dimensional Weyl-Richtmyer sequence. The mathematical development of the other low-discrepancy sequences[7] considered in this investigation, however, has

[6] In fact, the golden ratio $\phi = \frac{1}{2}(1 + \sqrt{5})$, or phi, has the continued fraction representation $\langle 1, 1, 1, \ldots \rangle$, which leads some (*e.g.* see [97]) to refer to phi as the most irrational number.

[7] These include: the Halton Sequence (Appendix C); the Faure sequence (Appendix E); the

been much more successful; and the implied constant in the asymptotic discrepancy bound is explicitly defined in terms of their constructive elements. As a result, it is possible to define an optimal set of constructive elements for generating each of these other low-discrepancy sequences such that the implied constant in their respective asymptotic discrepancy bounds $D_N^* = \mathcal{O}\left(N^{-1}(\log N)^s\right)$ is minimized. Note that more detail on the optimum sets of constructive elements for these sequences is given in Section 4.3 and Appendices C-F. The problem of finding the exact form of the implied constant $C(\mathbf{z})$ for the Weyl-Richtmyer sequence $S(\mathbf{z})$ is deeply rooted in the theory of Diophantine approximation, which has stymied progress since the asymptotic results of Niederreiter [123], Schmidt [155], and Zinterhof [194]. Niederreiter in [124] suggests that it may be possible to adapt the Jacobi-Perron algorithm [13] to develop an analogue to the one dimensional case, which identifies sets of irrational numbers that are poorly approximated by the rationals. There does not, however, appear any reference to this approach ever being successfully implemented in the literature review performed by the author. The modern development of practical applications for the QMC method has always been led by the mathematical progress in low-discrepancy sequences. As such, the absence of a known optimum set of irrational numbers for the Weyl-Richtmyer sequence may contribute, at least in part, to its current lack of popularity in the QMC literature.

The Weyl-Richtmyer sequence still holds some appeal due to the simplicity of its construction in spite of the problems encountered in its mathematical development. Without a rigorous mathematical result to guide the selection of the set of irrational numbers $\mathbf{z}$ used to generate the Weyl-Richtmyer sequence (except for the one dimensional result), the following engineering strategy is then proposed here. Note

Sobol' sequence (Appendix D); and the Niederreiter sequence (Appendix F).

in the construction of the multi-dimensional Weyl-Richtmyer sequence $S(\mathbf{z})$ in (4.3) that each dimension of the sequence is actually a one dimensional Weyl-Richtmyer sequence. Each irrational number in the set $\mathbf{z}$ is then chosen in such a manner as to achieve the lowest possible value for the constant $G(k) = \frac{2}{\log 2}$ in the discrepancy bound (4.5). This implies that all the irrational numbers in $\mathbf{z}$ have continued fraction representations in which all the coefficients are less than or equal to 3. When the irrational numbers in $\mathbf{z}$ are selected in this manner the Weyl-Richtmyer sequence is referred to as a BCF-3 (Bounded Continued Fraction) sequence in this investigation. The actual construction of $\mathbf{z}$ for the BCF-3 sequence is discussed in greater detail in Section 4.2.

In addition to the BCF-3 sequence, there are two other Weyl-Richtmyer sequences considered in this investigation: (i) the original implementation of Richtmyer in [148]; and (ii) the formulation discussed in the review paper of James [68]. In [148], Richtmyer constructs a 255 dimensional sequence using all the possible multiplicative combinations of the square roots of the first 8 prime numbers. More specifically, the set $\mathbf{z} = (z_1, \ldots, z_{255})$ is defined by

$$z_i = \left(\sqrt{2}\right)^{\xi_{8,2}(i)} \left(\sqrt{3}\right)^{\xi_{7,2}(i)} \cdots \left(\sqrt{19}\right)^{\xi_{1,2}(i)} \quad \text{for } 1 \leq i \leq 255, \tag{4.6}$$

where the vector $\vec{\xi}_2(i) = (\xi_{1,2}(i), \xi_{2,2}(i), \xi_{3,2}(i), \ldots)$ denotes the base 2 representation of the integer i as defined in (A.1). The somewhat curious construction of the set $\mathbf{z}$ in [148] was driven by the memory limitations of the computing machines in 1951, and was thus selected because only 8 irrational numbers would need to be stored. In [68], the i^{th} irrational number $z_i \in \mathbf{z}$ for the Weyl-Richtmyer sequence is defined simply as

$$z_i = \sqrt{p_i} \quad \text{for } n \geq 1, \tag{4.7}$$

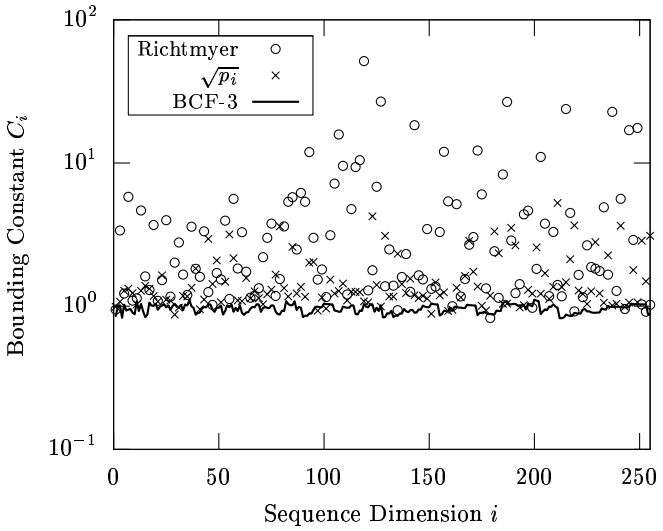

Figure 4.2: Comparison of the constant in the bounding inequality for the extreme discrepancy $D_N \leq C_i N^{-1} \log(N+1)$ for each dimension $1 \leq i \leq 255$ of different Weyl-Richtmyer sequences.

where p_i is the i^{th} smallest prime number.

By definition, the largest continued fraction coefficient found for the irrational numbers used to construct the BCF-3 sequence is 3. In comparison, the largest continued fraction coefficient present among the 255 irrational numbers defined for the original implementation of Richtmyer (4.6) is 6228 (for $z_{255} = \sqrt{9699690}$). Similarly, the largest continued fraction coefficient present among the first 255 irrational numbers defined for the formulation of James (4.6) is 80 (for $z_{255} = \sqrt{1613}$). To illustrate the actual effect of the continued fraction coefficients on the observed discrepancy bound $D_N(S(\mathbf{z}_i)) < C_i N^{-1} \log(N+1)$, the constant C_i is given in Figure 4.2 for the first 255 dimensions of each of the Weyl-Richtmyer sequences considered in this investigation. It should be noted that the results in Figure 4.2 are empirical by nature, and the value of C_i is determined to be the smallest value that satisfies the inequality $D_N(S(\mathbf{z}_i)) < C_i N^{-1} \log(N+1)$ for all sequence lengths $1 \leq N \leq 2^{16}$. Although lower than the actual theoretical upper bound given in (4.5), the empirical bounding con-

stants C_i are consistent with the expected behavior for these sequences. That is, the BCF-3 sequence typically has the lowest observed bounding constant, the sequence defined by James (4.6) has the next lowest bounding constant, and the sequence for the original implementation of Richtmyer (4.6) has the highest bounding constant.[8]

It is interesting to note that the Weyl-Richtmyer sequence is the only multi-dimensional low-discrepancy sequence that can be constructed in a manner such that each dimension is actually a one dimensional low-discrepancy sequence that achieves the lowest theoretical discrepancy bound. As the dimension of the other low-discrepancy sequences considered in this investigation increases, so too does the one dimensional discrepancy bound on the highest sequence dimension. This unique property of the Weyl-Richtmyer sequence is appealing; however, it does raise the question: "what is the benefit of having a good distribution of sample points in each dimension of a multi-dimensional integral approximation?" In the extreme case where each dimension of an s-dimensional function is completely separable under the integral operator (*e.g.* $f(\mathbf{x}) = x_1 + \cdots + x_s$), the approximation of the integral of $f(\mathbf{x})$ reduces to s one dimensional integrals. Of all the possible Weyl-Richtmyer sequences, the BCF-3 sequence yields the lowest theoretical error bound on the QMC integration error in this case due to the Koksma-Hlawka inequality (3.5) and the discrepancy bound in (4.5). Similarly, when the dimensions of a function are weakly-coupled under the integral operator, it is also expected that the BCF-3 sequence will yield a good approximation when used for the QMC integral approximation.

As presented in Section 5.5, each dimension of the low-discrepancy sequence is used to determine a new particle location within the free molecular duct for the

[8]For $1 \leq i \leq 255$, the following statistics on the bounding constant C_i are obtained for the three Weyl-Richtmyer sequences: (i) the original implementation of Richtmyer (4.6) – mean $C_i = 4.00$, min $C_i = 0.83$, max $C_i = 51.5$; (ii) the formulation of James (4.7) – mean $C_i = 1.54$, min $C_i = 0.87$, max $C_i = 5.66$; and (iii) the BCF-3 sequence – mean $C_i = 0.96$, min $C_i = 0.82$, max $C_i = 1.11$.

QMC particle simulation. While the trajectory angles generated for the new particle locations are physically independent of the previous trajectory angles,[9] the actual particle locations are not. In fact, the calculation of a next location for the particle is most affected by its current location. As the number of particle moves separating two trajectory locations increases, the impact of the earlier location on the future location diminishes. That is, the location of the particle after its 1^{st} move has a much greater affect on the location of the particle after its 2^{nd} than after its 10^{th} move. It is therefore reasonable to expect that some dimensions of the QMC particle simulation will be weakly-coupled. The arguments presented here are by no means rigorous; however, there does appear to be enough practical reasons to justify the development and testing of the BCF-3 sequence in this investigation for the QMC particle simulation.

4.2 Creating a Weyl-Richtmyer sequence with bounded continued fractions

To facilitate the discussion of the Weyl-Richtmyer sequence with bounded continued fractions, the concept of the BCF-k sequence is formally defined first.

Definition 4.3 *The s-dimensional* BCF-k sequence *is a Weyl-Richtmyer sequence $S(\mathbf{z})$ constructed from a set of irrational numbers $\mathbf{z} = \{z_1, \ldots, z_s\}$ such that the simple continued fraction of each irrational number $z_i \in \mathbf{z}$ (for $1 \leq i \leq s$) consists only of coefficients less than or equal to the integer k.*

As previously noted, there is some practical motivation for using a BCF-k sequence in a QMC particle simulations when the bounding constant k is small. The BCF-3 sequence, in particular, attains the lowest theoretical bound on the one dimensional

[9] This is a direct consequence of the assumption that the boundaries of the free molecular duct are fully-diffuse walls.

extreme discrepancy for each coordinate of the sequence. Definition 4.3 of the BCF-k sequences is merely categorical in nature; as such, it unfortunately does not offer any insight into the actual construction of the sequence.

The continued fraction representation of any of the irrational numbers used to construct the Weyl-Richtmyer sequence must contain an infinite number of coefficients (see Theorems 7.7 and 7.9 in [131]). Otherwise, if a number x has a continued fraction with a finite number of coefficients, then x is equivalently represented by a finite number of integer division operations. This would imply that x is a rational number. The construction of the BCF-k sequences thus requires the following two problems to be addressed: (i) how does one ensure that an infinite sequence of continued fraction coefficients remain bounded; and (ii) how does one ensure that the irrational numbers generated from the infinite continued fraction remain linearly independent over the rationals? Fortunately, there exists a simple method to solve both of these problems by constructing the BCF-3 sequence from quadratic surds.

A quadratic surd is an irrational number z of the form

$$z = \frac{b \pm c\sqrt{d}}{e},$$

where b, c, d, e are integers, and $d > 0$ and squarefree. In a theorem originally due to Lagrange (see Theorem 7.19 in [131]), the continued fraction of a quadratic surd is periodic. Stated more formally, if the continued fraction of a quadratic surd $z = \langle a_0, a_1, \ldots \rangle$, then for some integer $r \geq 0$ there exists an integer T such that the coefficients satisfy

$$a_{i+T} = a_i \quad \text{for all } i \geq r. \tag{4.8}$$

The period of the continued fraction for the quadratic surd is then defined as the smallest integer T that satisfies the condition in (4.8). For example, the continued

fraction of the quadratic surd $(6+\sqrt{2})/10$ is given by

$$\frac{6+\sqrt{2}}{10} = \langle 0, 1, 2, 1, 6, 1, 6, 1, 1, 6, 1, \ldots \rangle = \langle 0, 1, 2, \overline{1, 6, 1} \rangle,$$

where the pattern $\{1, 6, 1\}$ (period $T = 3$) repeats infinitely. Note that a vinculum, or over-line, is used throughout this investigation to denote an infinitely repeating pattern of the continued fraction coefficients. The advantage of the quadratic surds is that it is possible to determine the upper bound on the infinite sequence of continued fraction coefficients simply by checking a finite number of coefficients; specifically, the first $r + T$ coefficients (using the notation of (4.8)). It is possible to calculate the continued fraction of a quadratic surd using the procedure described by Knuth (see [78] p. 358) to generate the coefficients until the period is identified.[10] Once the complete periodic continued fraction is known, it is a simple matter to then provide an upper bound on its coefficients.

It is preferable, with respect to construction of the BCF-k sequences, to be able to convert a known periodic continued fraction with coefficients bounded by k to a closed-form representation of the irrational number. The alternative is to calculate the continued fractions of randomly selected quadratic surds using the procedure in [78] to check if the coefficients are bounded by k; this is an extremely inefficient method.[11] A general method for calculating an irrational number z directly from its continued fraction is obtained from the converging series of rational approximations to z. Given the continued fraction of an irrational number $z = \langle a_0, a_1, \ldots \rangle$, define $r_i = \langle a_0, a_1, \ldots, a_i \rangle$ as the i^{th} rational convergent to z for all $i \geq 0$. The i^{th} rational

[10] A word of caution is in order because the calculation of the continued fraction coefficients is extremely sensitive to round-off errors. Even under practical conditions using standard IEEE 64-bit double precision arithmetic, the procedure of Knuth [78] may become unstable after calculating only the first 10 coefficients.

[11] Except for a few special forms, it is difficult to determine the exact pattern of the continued fraction coefficients by inspection alone without actually performing the calculation procedure in [78].

convergent z is calculated by $r_i = p_i/q_i$, where p_i and q_i are defined by the following recursive formulae (see [127] p. 219),

$$\begin{aligned} p_{-2} &= 0, \quad p_{-1} = 1, \quad p_i = a_i p_{i-1} + p_{i-2} \quad \text{for } i \geq 0, \\ q_{-2} &= 1, \quad q_{-1} = 0, \quad q_i = a_i q_{i-1} + q_{i-2} \quad \text{for } i \geq 0. \end{aligned} \tag{4.9}$$

Note that the denominators $q_0, q_1, \ldots$ of the rational convergents monotonically increase (*i.e.* $1 = q_0 \leq q_1 < q_2 < \cdots$) because the continued fraction coefficients $a_1, a_2, \ldots$ are strictly positive by definition. More importantly, the i^{th} rational convergent to z is bound in the following equation (see Theorem 7.11 [131]),

$$|z - r_i| < \frac{1}{q_i q_{i+1}} \quad \text{for } i \geq 0,$$

implying, as the name already suggests, that the series $r_0, r_1, r_2, \ldots$ converges to the irrational number z. From the standpoint of generating the actual elements of the BCF-3 sequence for QMC integration, the series of rational convergents is an acceptable technique to calculate an irrational number from its continued fraction to the necessary accuracy on a finite precision machine. However, it is impossible to establish that a set of irrational numbers is linearly independent over the rationals by merely inspecting their respective continued fractions. The series of rational convergents defined by the recursive formulae in (4.9) is therefore an unsatisfactory method for producing the set of irrational numbers for the BCF-3 sequence because the necessary condition for establishing low-discrepancy (*i.e.* linear independence over $\mathbb{Q}$) cannot be proven.

While there may not exist a general method to determine the exact closed-form representation of any irrational number from its infinite continued fraction, there is a simple technique for converting a periodic continued fraction into its corresponding quadratic surd. To illustrate this technique, consider the simple case of determining

the quadratic surd z with a continued fraction $\langle\overline{1,3}\rangle$. Writing out the continued fraction yields

$$z = 1 + \cfrac{1}{3 + \cfrac{1}{\boxed{1 + \cfrac{1}{3 + \cfrac{1}{\cdots}}}}}.$$

Note that the region marked by the box is also a periodic continued fraction; in fact, it is the same continued fraction as z. Hence,

$$z = 1 + \cfrac{1}{3 + \cfrac{1}{z}}, \tag{4.10}$$

or $z = \langle 1, 3, z\rangle$. Simplifying the fraction terms in (4.10) yields a quadratic equation for z

$$3z^2 - 3z - 1 = 0,$$

with solutions $z = (3 \pm \sqrt{21})/6$. However, the solution $z = (3 - \sqrt{21})/6 < 0$ is not possible because all the terms in the continued fraction are positive. Therefore,

$$z = \langle\overline{1,3}\rangle = (3 + \sqrt{21})/6.$$

The process described for $z = \langle\overline{1,3}\rangle$ can be generalized to solve the exact closed-form of the quadratic surd represented by any periodic continued fraction. For purposes of numerical stability, however, only the quadratic surds that possess a purely periodic continued fraction are considered. Let $\langle\overline{a_0, \ldots, a_{T-1}}\rangle$ represent the purely periodic continued fraction of an irrational number z with a period T. As with the simple case, the irrational number z can be represented by an implicit

relationship using a general continued fraction with a finite number of coefficients; that is,

$$z = \langle \overline{a_0, \ldots, a_{T-1}, z} \rangle. \tag{4.11}$$

The recursive procedure defined in (4.9) for calculating the rational convergents is applicable to general continued fractions as well. Therefore, the implicit relationship (4.11) is simplified to yield

$$z = \frac{p_T}{q_T} = \frac{p_{T-1}z + p_{T-2}}{q_{T-1}z + q_{T-2}}, \tag{4.12}$$

where $p_0, \ldots, p_{T-1}$ and $q_0, \ldots, q_{T-1}$ are calculated from the continued fraction coefficients $a_0, \ldots, a_{T-1}$ using (4.9) as previously described. The implicit result in (4.12) can be rewritten to yield a quadratic equation for z

$$q_{T-1}z^2 - (p_{T-1} - q_{T-2})\, z - p_{T-2} = 0,$$

with a unique positive solution

$$z = \frac{(p_{T-1} - q_{T-2}) + \sqrt{(p_{T-1} - q_{T-2})^2 + 4p_{T-2}q_k}}{2q_{T-1}}. \tag{4.13}$$

It is important to note that the calculation in (4.9) for series of rational convergents $r_i = p_i/q_i$ for $i = 0, 1, \ldots$ becomes increasingly at risk to overflow errors as i increases. By only considering purely periodic continued fractions, this risk is alleviated to some extent by avoiding the additional calculation of the rational convergents associated with the non-repeating part of the continued fractions. To further reduce the risk of overflow errors, the set of irrational numbers used to construct the BCF-k sequences is calculated from continued fractions with the smallest possible period.

Now that a method is known to be capable of finding the closed-from quadratic surd associated with a purely periodic continued fraction, it is possible to detail how

the set of irrational numbers $\mathbf{z} = (z_1, \ldots, z_s)$ used to generate an s-dimensional BCF-k sequence $S(\mathbf{z})$ is constructed. Let Z_k denote an ordered infinite set of irrational numbers given by the following continued fractions,

$$\begin{aligned} Z_k = \big(&\langle\overline{1}\rangle, \langle\overline{2}\rangle, \ldots, \langle\overline{k}\rangle, \\ &\langle\overline{1,1}\rangle, \langle\overline{1,2}\rangle, \ldots, \langle\overline{1,k}\rangle, \\ &\langle\overline{2,1}\rangle, \langle\overline{2,2}\rangle, \ldots, \langle\overline{2,k}\rangle, \ldots \\ &\vdots \\ &\langle\overline{k,1}\rangle, \langle\overline{k,2}\rangle, \ldots, \langle\overline{k,k}\rangle, \\ &\langle\overline{1,1,1}\rangle, \langle\overline{1,1,2}\rangle, \ldots, \langle\overline{1,1,k}\rangle, \ldots\big) . \end{aligned} \tag{4.14}$$

By virtue of its design, Z_k defines an infinite set of purely periodic continued fractions that only contain coefficients less than or equal to k. Any linearly independent set of irrational numbers $\mathbf{z}$ over the rationals that is a subset of Z_k (4.14) can thus be used to construct a low-discrepancy BCF-k sequence.

As noted, there does not exist a general method to verify a set of irrational numbers is linearly independent over the rationals based solely on their continued fractions. It is possible, however, to quickly establish if a set of quadratic surds is linearly independent over the rationals using a very powerful theorem of Besicovitch. In [14], Besicovitch proves the necessary and sufficient conditions for a set of algebraic irrational numbers to be linearly independent over the rationals. Besicovitch's theorem, as it pertains to the construction of a BCF-k sequence, implies that a set of quadratic surds is linearly independent over the rational if and only if the squarefree integers appearing in the square roots of the quadratic surds are distinct. With the aid of Besicovitch's theorem, the set $\mathbf{z}$ of irrational numbers needed to construct a BCF-k sequence is then found by converting the continued fractions in Z_k (4.14)

to quadratic surds, and then simply eliminating from consideration any irrational numbers that are not linearly independent.

Let z_i' denote the irrational number corresponding to the i^{th} continued fraction in the ordered set Z_k in (4.14). Also, let D denote a set of squarefree integers used to track the linear independence of the quadratic surds. The set of irrational numbers $\mathbf{z} = (z_1, \ldots, z_s)$ used to generate an s-dimensional BCF-k sequence is then constructed using the following algorithm:

Algorithm 4.1

1. Initialize $i = 1$, $D = \emptyset$, and $\mathbf{z} = \emptyset$ (here $\emptyset$ denotes the null, or empty, set).

2. Find the quadratic surd $z_i' = (b_i \pm c_i\sqrt{d_i})/e_i$, where $b_i, c_i, d_i, e_i \in \mathbb{Z}$ and d_i is positive and squarefree, by converting the i^{th} continued fraction in Z_k (4.14) to a closed-form using equations (4.11-4.13).

3. If $d_i \notin D$, then add z_i' to the set $\mathbf{z}$ and d_i to the set D.

4. Increment i.

5. If $\text{card}\,(\mathbf{z}) < s$, then go to Step 2. Otherwise stop, as the set $\mathbf{z}$ now contains s irrational numbers with continued fraction coefficients bounded by k that are linearly independent over the rationals.

Note that the set of periodic continued fractions Z_k (4.14) contains many duplicates of the same irrational number. For instance, the continued fractions $\langle\overline{1}\rangle$,$\langle\overline{1,1}\rangle$, and $\langle\overline{1,1,1}\rangle$ clearly represent the same repeating pattern, and thus the same irrational number. The conversion from a periodic continued fraction to a quadratic surd using equations (4.11-4.13) yields the same result even if the period of the repeating coefficients is an integer multiple of the smallest period. Therefore, with respect to

Set $\mathbf{z}$	Continued Fraction	Quadratic Surd
z_1	$\langle\overline{1}\rangle$	$\frac{1+\sqrt{5}}{2}$
z_2	$\langle\overline{2}\rangle$	$1+\sqrt{2}$
z_3	$\langle\overline{3}\rangle$	$\frac{3+\sqrt{13}}{2}$
z_4	$\langle\overline{1,2}\rangle$	$\frac{1+\sqrt{3}}{2}$
z_5	$\langle\overline{1,3}\rangle$	$\frac{3+\sqrt{21}}{6}$
z_6	$\langle\overline{2,3}\rangle$	$\frac{3+\sqrt{15}}{3}$
z_7	$\langle\overline{1,1,2}\rangle$	$\frac{2+\sqrt{10}}{3}$
z_8	$\langle\overline{1,1,3}\rangle$	$\frac{3+\sqrt{17}}{4}$
z_9	$\langle\overline{1,2,2}\rangle$	$\frac{5+\sqrt{85}}{10}$
z_{10}	$\langle\overline{1,2,3}\rangle$	$\frac{4+\sqrt{37}}{7}$

Table 4.1: The first 10 irrational numbers used in the set $\mathbf{z}$ used to generate the BCF-3 low-discrepancy sequence.

actual implementation of the algorithm, it is much simpler to generate the entire set of continued fractions Z_k (4.14), and eliminate the duplicates during the linear independence check in Step 3 of Algorithm 4.1.

There is some impetus to construct the BCF-k sequence for use in QMC particle simulations; especially, as previously noted, in the case when $k = 3$. To illustrate the algorithm presented here for producing the set of irrational numbers $\mathbf{z}$ needed to generate a BCF-k sequence, the first 10 irrational numbers found by Algorithm 4.1 are given in Table 4.1 for the BCF-3 sequence. The s-dimensional BCF-3 sequence is generated from the ordered set of irrational numbers $\mathbf{z} = (z_1, \ldots, z_s)$, where z_i is the i^{th} quadratic surd added to the set $\mathbf{z}$ in Step 3 of Algorithm 4.1. Because the BCF-k sequence has not been the focus of any previous research into low-discrepancy sequences, there is no established standard choice for the set of irrational numbers used to construct the sequence. Thus, the construction of $\mathbf{z} = (z_1, \ldots, z_s)$ defined here is adopted throughout this investigation for the BCF-3 sequence.

The other low-discrepancy sequences used in the QMC particle simulations of this investigation (the Halton, Faure, and Niederreiter ($b = 2$) sequences) have a standard set of constructive elements that is used to generate the sequence.[12] These standard sets of constructive elements are typically chosen because they yield optimal performance in a certain theoretical sense. In particular, the Halton, Faure, and Niederreiter sequences have theoretical upper bounds on their star-discrepancy, which are minimized when generated from their standard constructive elements. While the actual asymptotic convergence rate for the star-discrepancy of these sequences $\mathcal{O}\left(N^{-1}(\log N)^{s-1}\right)$ is unchanged by the selection of the constructive elements; the implied constant of the convergence rate is affected. Without a similar theoretical upper bound on the star-discrepancy of a multi-dimensional Weyl-Richtmyer sequence, there is no obvious set of irrational numbers for generating the BCF-k sequence that should be made standard. However, there may be some practical advantage to be gained if the set of irrational numbers $\mathbf{z}$ for the BCF-k sequence is selected such that the correlation between any two dimensions of the sequence attains a minimum in some sense (see Section 6.4). Further investigation regarding the correlation of the BCF-k sequences is saved for future research.

For each of the low-discrepancy sequences tested in this investigation for the QMC particle simulations, the maximum required dimension of the sequences is 300. While there clearly exists an infinite number of continued fractions with coefficients bounded by a constant $k \geq 2$, it is not rigorously established if Algorithm 4.1 is capable of producing an infinite set of linearly independent irrational numbers for the BCF-k sequence. Let $\nu_k(T)$ denote the number of linearly independent quadratic

[12]The standard constructions of the s-dimensional low-discrepancy sequences are as follows: (i) the Halton sequence is constructed using the s smallest primes; (ii) the Faure sequence is constructed in the smallest prime base greater than or equal to s; and (iii) the Niederreiter sequences in base 2 is constructed from the s irreducible polynomials in $\mathbb{F}_2[x]$ with the smallest degree.

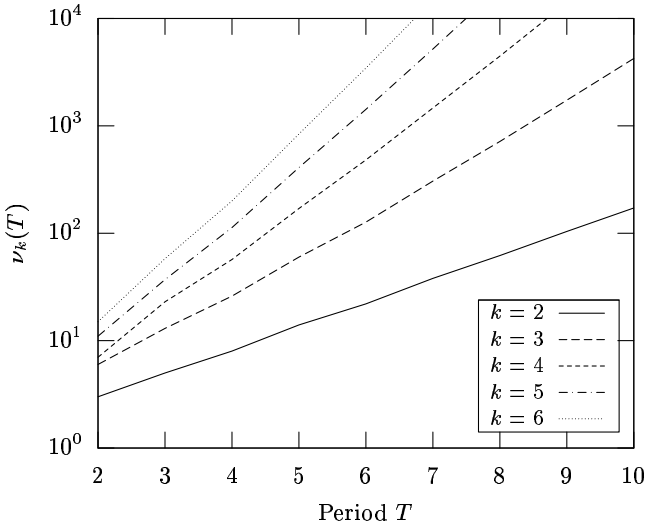

Figure 4.3: The number $\nu_k(T)$ of linearly independent quadratic surds which have a purely periodic continued fraction with coefficients bounded by k and a period less than or equal to T.

surds with a periodic continued fraction of the form $\langle\overline{a_0, \ldots, a_{m-1}}\rangle$, with $m \leq T$, and $a_i \leq k$ for $1 \leq i \leq m$. Because of the prescribed order of the set Z_k (4.14), the number $\nu_k(T)$ equals the maximum dimension of the BCF-k sequence that can be constructed using Algorithm 4.1 to search all periodic continued fraction patterns with a period less than or equal to T. Figure 4.3 illustrates the growth of $\nu_k(T)$ as the maximum period length T increases, for different bounding constants k. The number $\nu_k(T)$ appears to grow exponentially with the period length T; that is, $\nu_k(T) \propto T^{\alpha(k)}$, where $\alpha(k)$ only depends on the bounding constant k. The exponential growth of $\nu_k(T)$ in Figure 4.3 suggests that there is most likely an infinite number of irrational numbers with bounded continued fractions that are linearly independent over the rationals. No formal proof is considered here, however.[13] Most importantly, note that $\nu_3(7) = 306$, indicating that algorithm 4.1 is indeed capable of producing the

[13]Most likely, buried somewhere in the number theory literature of the 20^{th} century, there already exists a proof of the infinitude of linearly independent irrational numbers with bounded continued fractions. Unfortunately, the author is not able to locate such a proof at this time.

maximum required dimension for the BCF-3 sequence needed in the QMC particle simulations of this investigation.

4.3 Sequence Implementation

All the particle simulations for free molecular duct flow presented in this investigation are based on a sequence of vectors that are uniformly distributed throughout the unit hypercube $\overline{I}^s$; that is, $\mathbf{x}_1, \mathbf{x}_2, \ldots \in \mathscr{U}(0,1)^s$. Each element of the sequence $\mathbf{x}_n \in \overline{I}^s$ is used to generate up to s particle moves[14] for the n^{th} sample trajectory representing the particle behavior (see Section 5.5 for more details on the particle methods). Thus, it is important to understand the computational cost of generating the uniformly distributed sequence of vectors when assessing the computational performance of the particle simulation. There are five types of sequences that are employed in the various particle simulations presented here: the pseudo-random sequence; the Weyl-Richtmyer sequence; the Halton sequence; the Faure sequence; and the Niederreiter sequence in base 2. As reviewed in Appendices B-C,E-F, the four low-discrepancy sequences tested in this chapter have vastly different construction techniques; therefore, the computation time to implement the sequences is expected to vary considerably.

A comparison is given in Figure 4.4 of the computation time required to generate the five sequences used for particle simulations in this investigation with the sequence dimension s in the range $8 \leq s \leq 256$. For some of the sequences implemented here, the per-element cost of generating the sequence increases with the sequence length. In order to minimize the possible effect of sequence length, the results in Figure 4.4 are based on the computation time needed to generate the first $N = 2^{27}$ sequence

[14]The traditional test particle Monte Carlo method may not require all s dimensions to generate a sample trajectory because the particle may escape the duct in less than s moves.

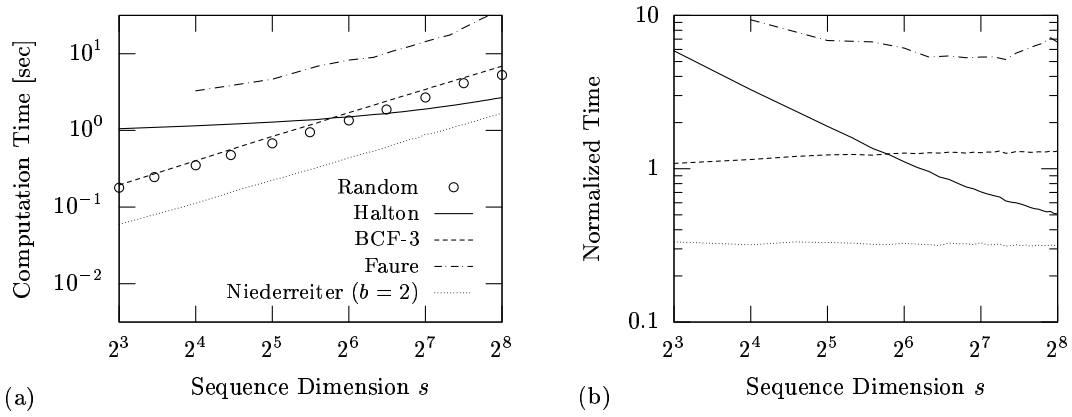

Figure 4.4: Comparison of the computation time needed to generate the pseudo-random sequence and the low-discrepancy sequences: (a) the computation time to generate a sequence of length $N = 10^6$; and (b) the computation time to generate the low-discrepancy sequence ($N = 10^6$) normalized by the time for the pseudo-random sequence.

elements.[15] The sequence length $N = 2^{27}$ corresponds to the maximum sequence length of the low-discrepancy sequence used in any of the QMC simulations in this chapter. It is important to note that the relative impact of the sequence generation on the total simulation time depends on the physical problem being simulated. Specifically, the importance of the sequence generation cost depends on the number of additional operations that must performed to transform a sequence element $\mathbf{x}_n$ into a sample representing the simulated stochastic process. For the particle simulations of the free molecular conductance probability in a duct, the cost of generating any sequence tested here contributes between 5% to 15% of the total simulation cost (except the Faure sequence). The high cost of generating the Faure sequence contributes more than 50% to the total simulation cost.

Before discussing the specific differences in the computational cost of the se-

[15]Except for the Faure sequence, which is based on the computation time to generate only the first $N = 2^{25}$ element. A smaller sequence length is used in this case because of the length computation time associated with generating the Faure sequence. Moreover, the per-element cost of generating the Faure sequence increases with the sequence length. Thus, the performance of the Faure sequence is actually 10% to 20% worse than the results in Figure 4.4.

quences used in the particle simulations, two additional points about their general implementation must be made. First, all the sequences are generated using the same hierarchy of abstract C++ class objects that is employed in the particle simulations presented in Sections 6.3 and 6.5. The multi-level hierarchy allows all the basic sequence operations needed by the actual simulation to be performed using the same base class; *e.g.* retrieving the components of a sequence element, and initiating the calculation of the next sequence element. Once the update is initiated, the specific construction details that are unique to the individual sequences are then performed in classes derived from the base class, which serves to hide their implementation from the main program. The benefit of the object-oriented program design is that the same particle simulation can be performed using any of the 5 sequences tested here without any changes to the main program despite the vast differences in actual sequence construction. However, the additional overhead associated with the hierarchy of abstract C++ class objects yields approximately 10% higher computation time than a dedicated sequence generation program. Second, the computation time for generating the sequences in Figure 4.4 includes a summation over all the components of each sequence. This summation is equivalent to a Monte Carlo, or QMC, approximation of the multidimensional integral,

$$\int_{\bar{I}^s} (u_1 + \cdots + u_s) d\mathbf{u} \approx \frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{s} x_{i,n}, \tag{4.15}$$

using the sequence $\mathbf{x}_n = (x_{1,n}, \ldots, x_{s,n})$ for $n = 1, 2, \ldots$. The simple summation ensures that each component of every sequence element generated is actually used at some later point during the execution of the program. This is important when performing timing studies because some compiler optimization techniques are robust enough to exclude certain calculations from the sequence generation if it is detected

that the sequence element is never used in the program. In some extreme cases, the compiler is clever enough to eliminate the entire sequence calculation yielding a computation time of zero. Note that the integral in (4.15) has an exact value of $s/2$ which can be used as a check value for the sequence generation.

Among the five sequences tested in the timing comparison in Figure 4.4, the generation of the Niederreiter sequence in base 2 is clearly the fastest. Specifically, the Niederreiter sequence in base 2 is three times faster than the pseudo-random sequence and the Weyl-Richtmyer sequence. The closest competitor is the Halton sequence in large dimensions, where the Niederreiter sequence remains at least 40% faster. The computation time for the Niederreiter sequence scales near linearly with the sequence dimension s, with almost no variation in cost with increasing sequence length. On the other end of the performance spectrum is the Faure sequence, which is 8-10 times slower than the pseudo-random sequence. The computation time of the Faure sequence also demonstrates a non-linear dependence on the sequence dimension and a per-element cost that increases with the sequence length. The computation time for the Halton, Weyl-Richtmyer, and pseudo-random sequences are all approximately on the same order. Both the generation of the Weyl-Richtmyer and pseudo-random sequences demonstrate nearly identical scaling with the sequence dimension, with the random sequence being slightly faster (between 10% to 30%); there is almost no variation in cost with increasing sequence length. The computation time for the Halton sequence has a strong dependence on the sequence dimension; however, it becomes relatively more efficient for longer sequences. For example, when the sequence dimension $s = 8$, the Halton sequence is 5 times slower than the pseudo-random sequence; yet, when the sequence dimension $s = 256$, the Halton sequence is actually 2 times faster than the pseudo-random sequence. Overall, the relative

computation times of the sequences tested in this investigation are consistent with earlier timing studies that appear in literature [48, 18, 116]. In order to understand the performance trends illustrated in Figure 4.4, the algorithmic implementation of each sequence tested in this investigation is briefly reviewed in the subsections that follow.

4.3.1 The pseudo-random sequence

The pseudo-random sequence is constructed using the pseudo-random number generator `random()` that is part of the standard Linux C package used by both the GNU C/C++ and the Intel C/C++ (version 8.0) compilers, for all the Monte Carlo simulations presented in this investigation. Quoting the Linux manual pages for `random()`, the function produces an unsigned 32-bit integer $z \in [0, 2^{31} - 1]$ using a "non-linear additive feedback random number generator." A general discussion of the design and implementation of additive feedback number generators is given by Knuth in [78], and the benefits of nonlinear generators are given in the review papers of L'Ecuyer [90, 91]. In order to obtain a pseudo-random sample from the uniform distribution $\mathscr{U}(0, 1)$, the integer z obtained from the function `random()` is scaled to the unit interval by multiplying z by the inverse maximum possible integer, `RAND_MAX` $= 2^{31} - 1$, that can be generated.

Generating a pseudo-random sequence in s dimensions simply requires s calls of the `random()` function and s multiplications to scale the sequence to I^s. Consequently, the computational time for the pseudo-random sequence implemented here scales linearly with the sequence dimension as illustrated in Figure 4.4. While the exact operation count of the `random()` function is not known by the author, a lower bound can be determined for comparison to the other low-discrepancy sequences

based on the implementation of the simpler additive feedback generator. The additive feedback generators with the lowest operation count are the lagged Fibonacci generators [78]. These generators are exceedingly fast and only require a single 32-bit addition, two references of non-sequential memory, and an increment of two memory pointers.

4.3.2 The Weyl-Richtmyer sequence

The mathematical description and asymptotic performance of the low-discrepancy Weyl-Richtmyer sequence is found in Appendix B. The construction of the Weyl-Richtmyer sequence in s dimensions requires a vector $\mathbf{z} = (z_1, \ldots, z_s)$ of irrational numbers to be selected that are linearly independent over the rationals. Let $\mathbf{x}_n = (x_{1,n}, \ldots, x_{s,n}) \in \overline{I}^s$ denote the n^{th} element of Weyl-Richtmyer sequence, then the sequence is defined using the vector $\mathbf{z}$ by

$$\mathbf{x}_n = ([nz_1], \ldots, [nz_s]),$$

where the square brackets $[\cdot]$ denote the fractional part of the argument, *i.e.* $[y] = y - \lfloor y \rfloor$. The generation of the Weyl-Richtmyer sequence can be recast in a more computationally efficient form by noting that

$$\mathbf{x}_n = ([z_1 + x_{1,n-1}], \ldots, [z_s + x_{s,n-1}]), \tag{4.16}$$

with $\mathbf{x}_0 = (0, \ldots, 0)$. Thus, it is possible to construct each element of the Weyl-Richtmyer sequence from the previous element with only s operations of addition and s operations to remove the integer part of the value.

In practice, it is not possible to calculate the exact Weyl-Richtmyer sequence because the irrational numbers in $\mathbf{z}$ can only be approximated using finite precision arithmetic. Therefore, one must select the working precision in which to perform the

operations in (4.16) in order to maintain a consistent level of accuracy throughout the construction of the sequence. In order to determine the appropriate working precision, assume that the irrational number z has the maximum possible truncation error of one half-bit when represented in this finite precision. As the truncation error propagates through the calculation in (4.16), the n^{th} element of the sequence calculation loses $\lfloor \log_2 n \rfloor$ bits of accuracy due to the initial half-bit error in z. To maintain an accuracy of m bits throughout the calculation of a Weyl-Richtmyer sequence with a maximum length N_{max}, the irrational numbers $\mathbf{z}$ must be stored and the calculation in (4.16) must be performed with $m + \lfloor \log_2 N_{max} \rfloor$ bits of precision. The maximum length restriction of any of the other sequences considered in this investigation is $N_{max} = 2^{32}$, which is adopted as a suitable limit for the implementation of the Weyl-Richtmyer sequence considered here. The calculations that generate samples for the conductance probability in the particle simulations are performed using 64-bit floating point arithmetic.[16] Moreover, the accuracy of the QMC simulation is so precise for certain duct geometries that the relative error is actually less than the machine error $\epsilon = 1.2 \cdot 10^{-7}$ of the 32-bit floating point arithmetic.[17] Therefore, it makes sense to adopt the requirement that the working precision of the Weyl-Richtmyer sequence calculation in (4.16) maintains accuracy comparable to the 64-bit floating point precision for a maximum sequence length $N_{max} = 2^{32}$.

A simple method for obtaining this level of accuracy uses 3 standard unsigned 32-bit integers to store a 96-bit representation of the irrational number z, and the n^{th} element x_n, for each dimension of the Weyl-Richtmyer sequence. Note that the subscripts for z and x_n, which denote the dimension of the sequence, are omitted in this part of the discussion for clarity. As a result of the $[\cdot]$ operation in (4.16),

[16] The IEEE standard 64-bit, double precision, floating point arithmetic.
[17] The IEEE standard 32-bit, single precision, floating point arithmetic.

the sequence element x_n remains in the unit interval $[0, 1)$ for all $n \geq 0$. Based on similar reasoning, the integer part of z does not affect the sequence calculation in (4.16); only $[z] \in [0, 1)$ must be stored. Since the quantities x_n and $[z]$ remain in the unit interval throughout the construction of the Weyl-Richtmyer sequence, it is possible to perform the extended precision calculation using fixed point arithmetic rather than floating point arithmetic. There are three advantages of fixed point arithmetic over floating point arithmetic: (i) it is much simpler to design extended precision algorithms using fixed point arithmetic; (ii) it is faster to perform the basic operations of fixed point arithmetic because there is no need to normalize the result after each intermediate step; and (iii) the operation $[\cdot]$ for obtaining the fractional part of the argument can be implemented at no cost by simply ignoring the addition overflow. One drawback to working in fixed point arithmetic is that values close to zero are unnormalized resulting in a minor loss of relative accuracy.

Using fixed point arithmetic scaled for the unit interval, the 96-bit representation for $[z]$ and x_n are formed by

$$[z] \approx 0.\underbrace{b_{96}\ \ldots\ b_{65}}_{\text{word } w_3}\underbrace{b_{64}\ \ldots\ b_{33}}_{\text{word } w_2}\underbrace{b_{32}\ \ldots\ b_{1}}_{\text{word } w_1}, \tag{4.17}$$

and

$$x_n \approx 0.\underbrace{b_{96}\ \ldots\ b_{65}}_{\text{word } v_{3,n}}\underbrace{b_{64}\ \ldots\ b_{33}}_{\text{word } v_{2,n}}\underbrace{b_{32}\ \ldots\ b_{1}}_{\text{word } v_{1,n}}, \tag{4.18}$$

where the computer words w_1, w_2, w_3 and $v_{1,n}, v_{2,n}, v_{3,n}$ are standard unsigned 32-bit integers. Adopting the notation from (4.17) and (4.18), the calculation of each dimension of the Weyl-Richtmyer sequence x_n is performed with 96-bit working precision

by the following algorithm:

$$
\begin{aligned}
v_{1,n} &= v_{1,n-1} + w_1, \\
v_{2,n} &= \begin{cases} v_{2,n-1} + w_2 + 1 & \text{if } v_{1,n} < w_1. \\ v_{2,n-1} + w_2 & \text{otherwise,} \end{cases} \\
v_{3,n} &= \begin{cases} v_{3,n-1} + w_3 + 1 & \text{if } v_{2,n} < w_2. \\ v_{3,n-1} + w_3 & \text{otherwise,} \end{cases} \\
x_n &\approx \frac{v_{3,n}}{2^{32}} + \frac{v_{2,n}}{2^{64}}.
\end{aligned} \tag{4.19}
$$

The algorithm (4.19) is essentially the same technique that is taught in primary school for basic addition, the only difference here is that the operations are performed in base 2^{32} instead of base 10. It is possible at some point during the calculation of $v_{i,n}$ (for $i = 1, 2, 3$), that the true value of $v_{i,n-1}+w_i$ is actually greater than 2^{32}. Referred to as an overflow calculation, the information that should have been represented by the 33^{rd} bit in this case is lost in the 32-bit representation of $v_{i,n}$. While some programming languages may allow direct access to an overflow flag associated with calculation; in general, it is possible to detect a calculation overflow by checking if $v_{i,n} < w_i$. A calculation overflow has occurred when $v_{i,n} < w_i$ is true and the 33^{rd} bit must be added to the next most significant word $v_{i+1,n}$; this process is also referred to as a carry operation. Any overflow that occurs for the most significant word $v_{3,n}$ that represents x_n can be ignored as it adds simply one to the integer value which is ignored by virtue of the $[\cdot]$ operation in (4.16). Thus, the construction of each dimension of each element of the Weyl-Richtmyer sequence using (4.19) is performed using 3 addition operations and 2 carry operations.

The minimum guaranteed accuracy of the first 2^{32} elements of the Weyl-Richtmyer sequence when generated by (4.19) is $\epsilon = 2^{-64}$, this error is absolute because the

calculations in (4.19) are performed using fixed point arithmetic. The 64-bit floating point representation of a number uses 52 bits to store the mantissa, excluding the implied bit due to normalization while the remaining 12 bits are reserved for the sign and exponent. Consequently, for the first 2^{32} sequence elements, there is an additional 12 bits of accuracy available when converting from the 96-bit fixed point representation to the 64-bit floating point representation. Thus, the conversion is performed without loss for all values of $x_n \in [2^{-12}, 1)$. There may be some loss in the conversion depending on the value of n, for values of $x_n < 2^{-12}$. Because the accuracy loss due to the accumulation of roundoff error in the calculation of x_n is $\lfloor \log_2 n \rfloor$, it implies that there is actually $44 - \lfloor \log_2 n \rfloor$ extra bits available for the conversion instead of 12. Therefore, any actual loss of accuracy during the conversion from 96-bit fixed point representation to the 64-bit floating point representation occurs rarely. If this implementation of the Weyl-Richtmyer sequence is applied to a QMC integration of a function with a singularity at zero, there may be some noticeable effect due to the fixed point calculations. However, the presence of such a singularity would make any application of the QMC method suspect. For the QMC simulation of the conductance probability in a free molecular flow, the resulting integrand is smooth and bounded, and any effect due to the lack of floating point arithmetic is negligible at best.

With the algorithm (4.19) for constructing the Weyl-Richtmyer sequence in place, the focus now shifts to the selection of the set $\mathbf{z}$ of s irrational numbers used to generate the sequence. Unlike the other low-discrepancy sequences, there does not exist any mathematical theory to guide the selection of $\mathbf{z}$ for the multi-dimensional Weyl-Richtmyer sequence. For the Halton, Sobol', Faure, and Niederreiter sequences, it is

possible to choose the constructive elements[18] such that the asymptotic constant in the discrepancy bound for the sequence is minimized. Except in the one dimensional case, there is not a comparable result to exploit for the Weyl-Richtmyer sequence. Based on the heuristic arguments in Section 4.1, there is some motivation to select $\mathbf{z}$ from the set of irrational numbers with continued fraction coefficients that are all bounded by a small constant integer. Such a sequence is referred to as a BCF-k sequence in this investigation and is defined as follows. A Weyl-Richtmyer sequence is a BCF-k sequence if all the continued fraction coefficients of the irrational numbers in $\mathbf{z}$ used to construct the sequence are less than or equal to the integer k. Please refer to [75, 131] for a thorough review of the theory of continued fractions. Based on the theoretical results [127] and the empirical results (see Figure 4.2), when the bounding constant k on the continued fraction coefficients is smaller for a given irrational number z, there is also a smaller constant in the extreme discrepancy bound (4.5) for the sequence constructed from z. In fact, the theoretical bound in (4.5) is the smallest when $k = 1, 2, 3$. Therefore, the BCF-3 sequence is of practical interest here for the QMC particle simulation because, at the very least, the sequence is expected to yield a good QMC approximation for problems where the dimensions are weakly-coupled under the integral operator.

To find the best representative of a Weyl-Richtmyer sequence for comparison to the other low-discrepancy sequences, the same three sets of irrational numbers tested in Figure 4.2 are used for the QMC particle simulations of free molecular duct flow. Specifically, these include the following choices of the irrational set $\mathbf{z}$: (i) the original implementation of Richtmyer [148] using combinations of the square roots of the

[18]The term "constructive elements" used here refers to the prime bases of the Halton and Faure sequences, the primitive polynomials of the Sobol' sequence, and the irreducible polynomials of the Niederreiter sequence.

first 8 prime numbers, as defined in (4.6); (ii) the formulation described in the review paper of James [68] using the square roots of the smallest prime numbers, as defined in (4.7); and (iii) the BCF-3 sequence using the procedure outlined in Section 4.2. It is important to note that the actual development of the QMC particle simulation appears later in Section 5.5. However, it is necessary to present some results of the QMC particle simulation here in order to establish which Weyl-Richtmyer sequence yields the best performance. The results given in this section for the particle simulations are, therefore, only discussed in terms of the general error convergence without any mention to the specific implementation details of the QMC method.

The performance of the QMC particle simulation for approximating the conductance probability for free molecular duct flow is given in Figure 4.5 for a duct length to height ratio L in the range of $0.5 \leq L \leq 9.5$, using the three Weyl-Richtmyer sequences. To reduce the noise in the error convergence results in Figure 4.5, 512 ensembles of the traditional test particle Monte Carlo simulation are collected and then averaged together. Similarly, 16 ensembles of the QMC particle simulation are also averaged together.[19] Compared to the other two classic Weyl-Richtmyer sequences, the BCF-3 sequence produces the lowest relative error at $N = 2^{23}$ for all the duct geometries under consideration (see Figure 4.5(a)). In particular, the BCF-3 sequence yields an error that is significantly smaller than the traditional test particle Monte Carlo simulation by a factor of 10^4 when $L = 9.5$, and by a factor of 12 when $L = 9.5$. Furthermore, for most of the duct geometries tested, the BCF-3 sequence produces an error that is 1.5 to 3 times smaller than the other two classic implementations of the Weyl-Richtmyer sequence.

The error convergence rate, for each particle simulation (Monte Carlo and QMC),

[19]The 16 ensembles for the QMC simulation are constructed from 16 equal length subsequences of the same low-discrepancy sequence.

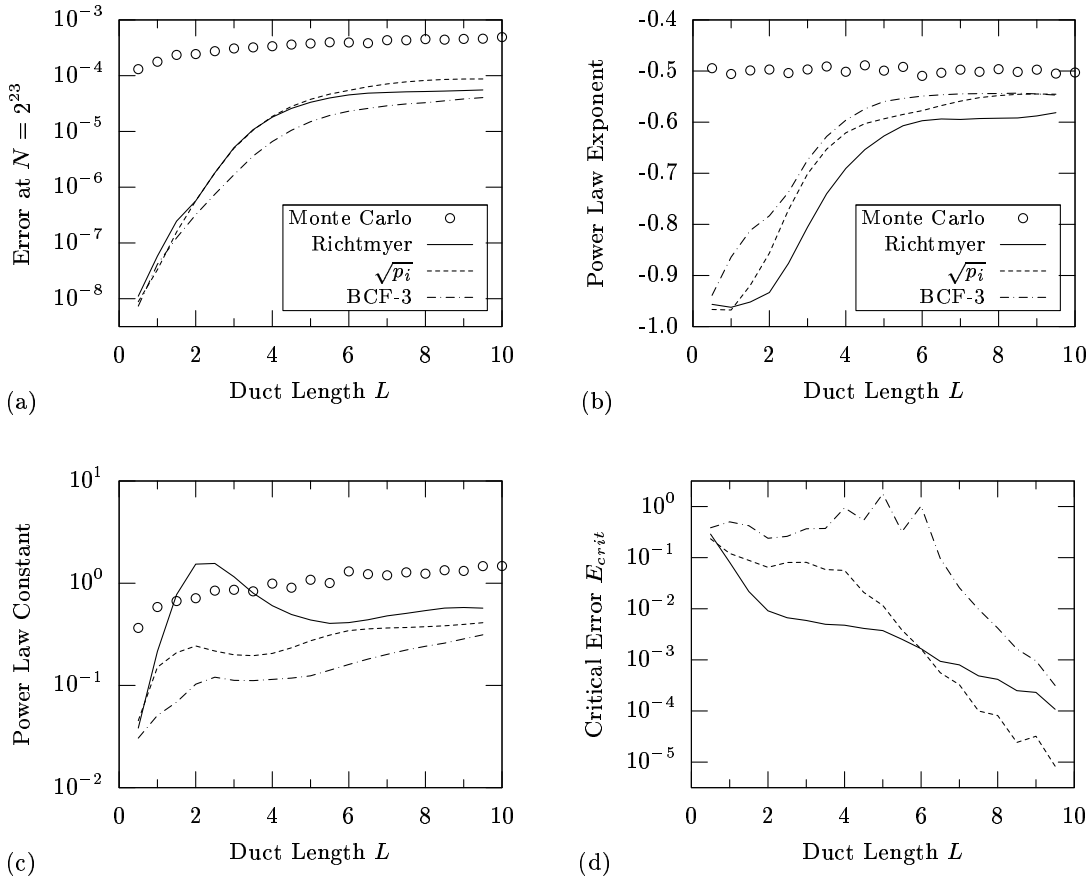

Figure 4.5: QMC performance comparison using different Weyl-Richtmyer low-discrepancy sequences: (a) the relative error after $N = 2^{23}$ samples; (b) the power law exponent for the error convergence rate; (c) the power law constant for the error convergence rate; and (d) the maximum error at which the QMC simulation remains faster than DSMC.

is modeled as a power law; that is, the error $E = cN^{\gamma}$, where N is the number of samples. The power law exponent γ and the power law constant c are given in Figure 4.5(b) and (c) respectively for each of the duct geometries tested. Interestingly, the BCF-3 sequence has the slowest error convergence rate (*i.e.* the power law exponent γ is the least negative) of all the Weyl-Richtmyer sequences. The BCF-3 sequence is still able to achieve the lowest error after $N = 2^{23}$ samples, in spite of its slower convergence rate, because its power law constant c is significantly smaller than the other sequences (see Figure 4.5(c)). More specifically, the power law constant for the convergence of the BCF-3 sequence is 1.5 to 15 times smaller than the other two Weyl-Richtmyer sequences when the duct length to height ratio L is in the range $1 \leq L \leq 8.5$. This performance observed for the BCF-3 sequence is to be expected from its intended design. Each dimension of the BCF-3 sequence, after all, is designed to have the lowest possible bounding constant for the convergence of the one dimensional discrepancy. The BCF-3 sequence should then, as a consequence of the Koksma-Hlawka inequality, also possess the lowest bounding constant on the error convergence in QMC simulations where the problem dimensions are weakly-coupled under the integral operator.

The power law models of the error convergence for the Monte Carlo and QMC particle simulation can be combined with the overall computation time to determine a critical relative error E_{crit} level that serves as a demarcation between the performance of the two methods. Recall from Figure 4.5(b) that the convergence rate of the QMC methods is greater than the rate of the Monte Carlo method $\mathcal{O}\left(N^{-1/2}\right)$ for all the Weyl-Richtmyer sequences and duct geometries tested. Consequently, the simulation time to reach any error level below E_{crit} is faster for the QMC particle simulation than the Monte Carlo method (*i.e.* the computation time $\tau_{qmc} < \tau_{mc}$). A larger value

for the critical error E_{crit}, therefore, indicates that the QMC simulation is the more computationally efficient method for a wider range of desired simulation accuracies. As shown in Figure 4.5(d), the BCF-3 sequence has the largest critical error E_{crit} (where $\tau_{qmc} < \tau_{mc}$) by virtue of having the smallest power law constant for the error convergence. In fact, for duct geometries with $L \leq 6.5$, the QMC simulation with the BCF-3 sequence is faster than the Monte Carlo method for achieving any relative error less than 10%, which covers most simulation accuracy levels of practical interest. Thus, the BCF-3 sequence developed here is used for the QMC simulations as the representative low-discrepancy Weyl-Richtmyer sequence for the remainder of the investigation.

The cost of implementing the different Weyl-Richtmyer sequences is not a factor because the choice of irrational numbers $\mathbf{z}$ used in the construction does not affect the calculation in (4.16). The generation of the BCF-3 sequence requires three addition operations and two carries to be performed for each dimension of each sequence element, and each dimension is constructed in the same manner. Hence, the cost of generating the BCF-3 sequence is expected to be linear with sequence dimension s. There is, however, a slight increase in computation time with s when compared to the random sequence that has a true linear dependence between computation time and sequence dimensions, as shown in Figure 4.4(b). The computation time for the BCF-3 sequence is 10% slower than the random sequence when $s = 8$, and 30% slower when $s = 256$. The slight non-linear increase in computational cost with dimension is most likely attributed to the increased amount of computer memory that must be accessed by the algorithm in (4.16). These types of performance effects depend on the relative size and architecture of the L1 and L2 memory caches on the actual computer chip, and the compiler options used to create an executable version of the

code.

4.3.3 The Halton sequence

The mathematical description and asymptotic performance of the low-discrepancy Halton sequence is found in Appendix C. The i^{th} dimension of Halton sequence used in this investigation is generated from a one dimensional van der Corput sequence in base p_i, where p_i is the i^{th} smallest prime number for $1 \leq i \leq s$.[20] It is often convenient to describe the base b van der Corput sequence in terms of the inverse radical function $\chi_b(n)$ (see Appendix A) Specifically, the sequence $(\chi_b(0), \chi_b(1), \chi_b(2), \ldots)$ represents the van der Corput sequence in base b. The base b representation function $\vec{\xi}(n)$ in (A.1) can be combined with the definition of $\chi_b(n)$ in (A.2) to yield a recursive form for $\chi_b(n)$ given by

$$\chi_b(n) = \begin{cases} \frac{1}{b}\big(\chi_b(\lfloor \frac{n}{b} \rfloor) + \mathrm{mod}(n, b)\big) & \text{if } n > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{4.20}$$

The recursive form in (4.20) is appealing because it can be implemented as a single-line function in most programming languages; however, it is exceedingly wasteful. The calculation of each recursive step of $\chi_b(n)$ requires one addition, multiplication, division, and modulo operation. Furthermore, the number of steps m until the recursion terminates for the n^{th} element of the van der Corput sequence is equal to the number of digits in the base b representation of n; *i.e.* $m = \lfloor \log_b n \rfloor + 1$.

In order to understand why most of the operations performed in the recursive form (4.20) are unnecessary, consider the following calculation of $(b-1)$ consecutive elements of a base b van der Corput sequence. Let k denote any non-negative integer, and suppose the $(bk)^{th}$ element of the van der Corput sequence $\chi_b(bk)$ is known. Then

[20] Except where otherwise noted in Section 6.4.

the next $(b-1)$ elements of the sequence are found from the more computationally efficient recursion,

$$\chi_b(bk+r) = \chi_b(bk+r-1) + \frac{1}{b}, \text{ for } r = 1, 2, \ldots, b-1. \tag{4.21}$$

At the very worst, one need only perform the direct calculation of the inverse radical function $\chi_b(n)$ using (4.20) (or any other means) when $n \equiv 0 \pmod{b}$, only occurring for $(\frac{1}{b})^{th}$ of the sequence elements. The remaining fraction of the elements (equal to $\frac{b-1}{b}$) may be calculated by a single addition of the constant $\frac{1}{b}$ to the previous sequence element. This simplification offers a tremendous decrease in computation time over the direct calculation of the inverse radical function for every element of the van der Corput sequence. Moreover, the average cost of computing the van der Corput sequence in base b decreases as b increases.

In the original description of Halton sequence in [56], Halton provides a very efficient algorithm for constructing the van der Corput sequences based on the additive recursion in (4.21). Furthermore, the algorithm in [56] does not perform the full recursion in (4.20) for the n^{th} sequence element when $n \equiv 0 \pmod{b}$. In this case, the algorithm finds $\chi_b(n)$ using the previous element $\chi_b(n)$ in α steps, where α is the largest integer such that $n \equiv 0 \pmod{b^\alpha}$. While the algorithm in [56] is extremely fast, it is not generally stable using finite precision arithmetic. In [57], Halton and Smith discuss general conditions for preventing the onset of an unstable sequence calculation and they outline a modification to the original algorithm in [56] to avoid one type of instability without affecting the computation time. In comparison to the other sequences implemented in this investigation, the modified algorithm in [57] generates the Halton sequence in the second fastest overall computation time; only the Niederreiter sequence in base 2 is generated faster. Typically, the generation of

Niederreiter sequence in base 2 is approximately 30-35% faster than the modified Halton algorithm for the sequence dimensions and lengths under consideration here.

Initially in this investigation the Halton sequence was implemented using the modified algorithm outlined in [57]. However, a numerical instability was eventually discovered in the generation of the higher dimensions of the sequence; specifically when the prime bases are greater than 1,000. When generating a dimension of the sequence in base b, the instability would occur sometimes for the calculation of the n^{th} element when $n \equiv 0 \pmod{b}$. The floating point calculations are performed too close to the machine precision, in these cases of instability. It is not clear to the author at this time how to prevent the stability problem in all the dimensions of the sequence; and the details of the modifications suggested in [57] are only briefly sketched. Fox in [48] also follows the modified algorithm of Halton and Smith [57]; moreover, Fox provides an explicit method to check the anticipated stability of the sequence generated. Unfortunately, the stability check in [48] appears only in the source code accompanying the journal article. Instead of updating a 2 decade old FORTRAN77 code and verifying is reliability for the sequence dimensions needed here, the simple but costly full recursion definition in (4.20) is used as necessary to maintain a stable calculation.

The numerical instability is only present for the calculation of the n^{th} element of the Halton sequence, when n is divisible by the base b of the associated van der Corput sequence. In contrast, the additive recursion in (4.21) that is used to construct the $(b-1)$ consecutive elements is always a stable operation. Consequently, to achieve a stable construction of the Halton sequence, a stable calculation of $\chi_b(n)$ is needed when $n \equiv 0 \pmod{b}$. In the initial implementation of the Halton sequence, the instability occurs because of the presence of truncation errors in the floating point

calculations. Alternatively, the full recursion in (4.20) avoids such truncation errors by treating the division and floor operations in the argument of $\chi_b(\lfloor n/b \rfloor)$ as integer calculations. Thus, the final implementation of the Halton sequence used in this investigation may be stated as follows. Let $\mathbf{x}_n$ denote the n^{th} element of a Halton sequence in s dimensions; that is, $\mathbf{x}_n = \big(\chi_{p_1}(n), \chi_{p_2}(n), \ldots, \chi_{p_s}(n)\big)$, where p_i is the i^{th} smallest prime for $1 \leq i \leq s$.[21] Then for each sequence dimension $1 \leq i \leq s$, the inverse radical function $\chi_{p_i}(n)$ is calculated using the full recursion in (4.20), only if $n \equiv 0 \pmod{p_i}$; otherwise, $\chi_{p_i}(n)$ is calculated using the additive recursion in (4.21).

The resulting method is slower than the modified Halton method in [57], especially for sequences with few dimensions constructed in small prime bases. While it is generally appealing to use the most computationally efficient methods to generate the sequences, the performance loss in the final implementation of the Halton sequence is not a primary concern of this investigation. Even with the most efficient - albeit unstable implementation - the Halton sequence is slower than the Niederreiter sequence in base 2. In fact, it is the Niederreiter sequence in base 2 that is shown in Section 6.3 to offer the best error convergence, irrespective of computation time, for nearly all of the duct geometries simulated in this chapter. Therefore, even with the best possible algorithmic implementation for the Halton sequence, it still would not serve as the best representative of the QMC particle simulations for comparisons with Monte Carlo.

As the dimension of the Halton sequence increases, so too does the prime base used in the construction of the van der Corput sequences increase for each dimension. Thus, the fraction of sequence components calculated with the additive recursion in (4.21) also increases with the sequence dimension. Let η denote the fraction of

[21] Except where otherwise noted in Section 6.4.

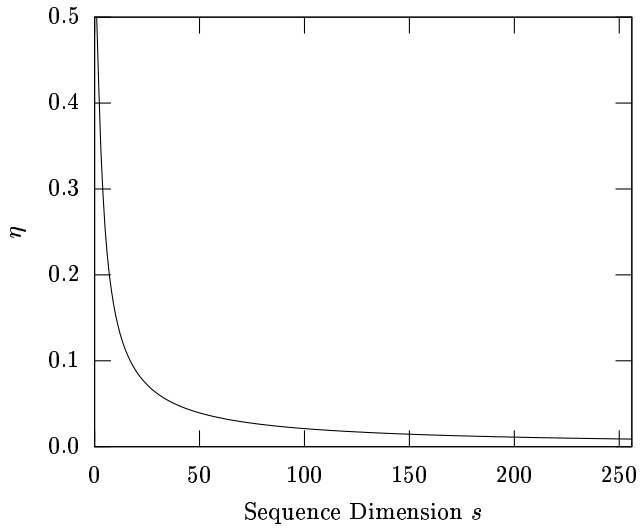

Figure 4.6: The fraction of the components η of Halton sequence in s dimensions that can be calculated using the simple additive recursion.

sequence components calculated with the additive recursion, then the dependence of η on the sequence dimension is illustrated in Figure 4.6. The combined effect of the initial rapid decrease in η with the sequence dimension, and the tremendous cost savings attributed to the additive recursion, results in only a 50% increase in computation time when the sequence dimension is increased 8-fold from $s = 8$ to $s = 64$. Consequently, the computation time for the Halton sequence does not increase linearly with sequence dimension s. In fact, when compared to the random sequence that has a construction time linearly increasing with dimension, the Halton sequence, as implemented here, is 5 times slower for $s = 8$ and 3 times faster for $s = 256$, as shown in Figure 4.4.

4.3.4 The Faure sequence

The mathematical description and asymptotic performance of the low-discrepancy Faure sequence is found in Appendix E. The construction of the Faure sequence in s dimensions requires a prime base $q \geq s$ to be selected in which all the operations are

performed. In order to minimize the constant in the asymptotic discrepancy bound in (E.1), q is selected to be the smallest prime number greater than or equal to s.[22] The implementation of the Faure sequence in this investigation follows in part the algorithm of Fox [48]. The calculation of each dimension of each sequence element begins with a matrix-vector multiplication using the binomial coefficient matrix C defined in (E.2). For each dimension, the vector result of the matrix-vector multiplication is then mapped to the final value of the sequence element in the unit interval $[0, 1)$ by taking its dot product with the constant vector $(q^{-1}, q^{-2}, \ldots, q^{-k-1})$, where the maximum sequence length considered is less than q^{k+1}.

Faure [46] originally defines the construction of his sequence in terms of the powers of the binomial coefficient matrix; that is, $C, C^2, C^3, \ldots, C^{q-1}$. However, in an effort to reduce the memory overhead as suggested by Fox in [48], only the matrix C is pre-computed and stored in the implementation of this investigation. The remaining matrix powers are computed implicitly using the recursive definition of the matrix-vector operation given in (E.3). It is important to note that the matrix-vector multiplication using C is performed over the finite field $\mathbb{F}_q$, which requires modular arithmetic. A key difference between the implementation presented here for the Faure sequence and the algorithm proposed by Fox in [48] is how these field operations are performed. The finite field calculations, in this investigation, are performed using pre-computed tables for the addition and multiplication operations over $\mathbb{F}_q$. In contrast, the algorithm of Fox uses the standard definitions of addition and multiplication for the matrix-vector multiplication, and then performs a modulo operation in base q for each row of the matrix. The savings in computational time associated with the pre-computed field operation tables is 15% to 20%, where the

[22] Except where otherwise noted in Section 6.4.

higher savings occurs when the Faure base q is small, or the sequence length is long. The use of pre-computed field operation tables is not a new idea to the construction of low-discrepancy sequences, and has been employed by Bratley *et. al.* in [19] for the generation of the general Niederreiter sequence in bases $b > 2$.

The majority of the computational cost of the Faure sequence is due to the matrix-vector multiplication. In order to calculate the n^{th} element of the Faure sequence, the minimum required size of matrix C must be at least $m \times m$, where m is the number of digits in base q needed to represent n (*i.e.* $m = \lfloor \log_q n \rfloor + 1$). Some savings can be achieved in this step by exploiting the fact that the matrix C is upper triangular, and the entire matrix is not needed to compute every sequence element. In this case, the matrix-vector multiplication is performed with $\frac{1}{2}m(m+1)$ addition and multiplication operations over $\mathbb{F}_q$. The result of the matrix-vector multiplication is then mapped to the unit interval for the final value of the sequence element with m addition and multiplication operations. Therefore, the total operation count is $\mathcal{O}(m^2)$ for generating each dimension of each element of the Faure sequence. Note that the minimum size m of the matrix C increases with the sequence length; thus, the per-element computational cost also increases with the sequence length.

In spite of the various attempts at minimizing the computational costs, the total number of operations necessary to construct the Faure sequence far exceeds any other sequence implemented in this investigation.[23] Consequently, the time required to generate the Faure sequence is approximately 8-10 times greater than any other in this investigation (see Figure 4.4). Interestingly, the minimum required size of the matrix C is generally larger when the Faure base q is smaller for the same sequence

[23]The implementation of the general Niederreiter sequence in a base greater than 2 also requires a similar matrix-vector multiplication over a finite field. As such, the general Niederreiter sequence requires approximately the same computation time as the Faure sequence. However, the sequence is not considered for the QMC particle simulations in this investigation.

length. The increase in the computation cost is thus partially offset as the sequence dimension s increases (and hence q) because the matrix C has fewer elements. As a consequence, the slope of the computation time versus sequence dimension in Figure 4.4 is initially less than the sequences with a linear dependence on dimension (*e.g.* the pseudo-random sequence).

4.3.5 The Niederreiter sequence in base 2

The mathematical description and asymptotic performance of the low-discrepancy Niederreiter sequence is found in Appendix F. The s-dimensional Niederreiter sequence in base 2 is constructed from s distinct irreducible polynomials over the finite field $\mathbb{F}_2$ with the smallest possible degree.[24] Each dimension of the Niederreiter sequence uses one of the distinct irreducible polynomials in $\mathbb{F}_2[x]$ to construct the matrix transform A defined in (F.4). The order of the irreducible polynomials is chosen such that for any sequence dimension k, the degree of the polynomial used to construct the k^{th} dimension is less than or equal to the degree of the polynomial used to construct the $(k+1)^{th}$ dimension. Because of the enumeration technique adopted in this investigation to generate the irreducible polynomials in $\mathbb{F}_2[x]$, the exact order of the polynomials is the same as the table given in the appendix of [96].

The implementation of the Niederreiter sequence in base 2 in this investigation closely follows the algorithm developed by Bratley *et. al.* in [19]. Their algorithm exploits two key features of the base 2 Niederreiter sequence in order to produce a very computationally efficient method for generating the sequence. In addition to these performance enhancements, the algorithm in [19] uses a leading zero correction to the components in the A matrix (F.4) used to generate each dimension. The leading zero correction reduces the correlation problems present between the dimensions of

[24] Except where otherwise noted in Section 6.4.

the sequence at startup, and is adopted in this investigation as well.

The general construction of the Niederreiter sequence in an arbitrary base q, where q is a prime power, requires a matrix-vector multiplication over $\mathbb{F}_q$ similar to the Faure sequence. The first key feature that Bratley *et. al.* exploit is that in base 2 the elements of the matrix A (F.4) used in the matrix-vector multiplication are simply ones and zeros. It is therefore possible to represent an entire row of the matrix A as a single computer word, using each bit to represent an element of the matrix. In doing so, the matrix-vector operation for the Niederreiter sequence in base 2 is reduced to an operation equivalent to a vector dot product. In this investigation the computer word is chosen to be the IEEE standard 32-bit unsigned long integer that is used by the C/C++ compilers on the Linux platform. Therefore, the maximum length of the Niederreiter sequence in base 2 that can be constructed here is $N = 2^{32}$, which is sufficient for all the QMC simulations performed in this investigation.

The second key feature that Bratley *et. al.* exploit is that the binary Gray code can be used to represent the sequence order for the Niederreiter sequence in base 2.[25] Using the binary gray code to change the sequence order was first proposed by Antonov and Saleev in [2] for the low-discrepancy Sobol' sequence. The binary gray code modification in [2] essentially permutes the original order of the Sobol' sequence in blocks of 2^i elements, for $i = 1, 2, \ldots$, without affecting the asymptotic bounds on its star discrepancy. Note that the construction of the n^{th} element of Sobol' sequence requires the binary representation of the n for the calculation (see Appendix D). For each non-zero bit in the binary representation of the n, a single bit-wise XOR operation is performed using a set of constant computer words referred to as "direction numbers" by Sobol'. The results of these bit-wise XOR operations

[25]The binary Gray code is linked to many classic mathematical puzzles, such as the Towers of Hanoi and the baguenaudier ring puzzles [52].

are then accumulated for each non-zero bit to generate the actual element of the Sobol' sequence. The advantage of using the binary gray code to represent the sequence order n is that there is only one bit difference between consecutive values of n. Therefore, using the binary Gray code, each element of the Sobol' sequence is calculated from the previous element of the sequence using a single bit-wise XOR operation for each dimension. In contrast, the original construction of the n^{th} element of the Sobol' sequence requires up to $\lfloor \log_2 n \rfloor + 1$ bit-wise XOR operations for each dimension. Moreover, each of the bits of n must be checked for non-zero values to determine the number of bit-wise XOR operations that are performed, which further adds to the computational cost of the original construction of the Sobol' sequence. For the binary Gray code modification to the Sobol' sequence, Antonov and Saleev report in [2] a reduction in computation time by a factor of 5.6 over the original design for generating $5 \cdot 10^5$ sequence elements. Furthermore, the computational cost associated with the binary Gray code modification increases with the sequence length due to the need to check more bits of the sequence order number n.

The Niederreiter sequence in base 2 is nearly identical to the Sobol' sequence. The constant "direction numbers" used in the construction of the Sobol' sequence are equivalent to the columns of the matrix A (F.4) used in the construction of the Niederreiter sequence in base 2. Similarly, the accumulation of the bit-wise XOR operations for the Sobol' sequence is equivalent to the general matrix-vector multiplication for the Niederreiter sequence in base 2. Bratley *et. al.* in [19] exploit these similarities between the two sequences to apply the binary Gray code modification of Antonov and Saleev [2] in the same manner to the Niederreiter sequence in base 2. The resulting algorithm in [19] achieves a similar computational speedup as in [2] because the entire matrix-vector multiplication is replaced by a single bit-wise

XOR operation in each dimension. Therefore, because of the apparent performance enhancements, the implementation of the Niederreiter sequence in base 2 in this investigation follows closely the algorithm design of Bratley *et. al.* in [19].

The generation of each sequence element $\mathbf{x}_n$ of the Niederreiter sequence in base 2 requires the determination of the Gray bit of n. In the Gray code representation of n, the Gray bit is the bit that changes parity from $n-1$ to n. Hence, the Gray bit indicates which column of the matrix A (F.4) is used in the bit-wise XOR operation for each dimension to update the sequence from $\mathbf{x}_{n-1}$ to $\mathbf{x}_n$. The Gray bit of an integer n corresponds to the location of the least significant zero in the standard binary representation of n, making it relatively simple to determine. The update of the s dimensional Niederreiter sequence in base 2 is performed with s bit-wise XOR operations, once the Gray bit is known. The bit-wise XOR operation is comparable to a single addition operation, although generally faster because there is no need to perform a carry operation for each bit. In addition to the bit-wise XOR operation, a multiplication is also needed for each dimension to scale the result to the unit interval in the same manner as the random, Weyl-Richtmyer, and Faure sequences. Note that the search for the Gray bit is only performed once for each sequence element $\mathbf{x}_n$, regardless of dimension. While the implementation of the Niederreiter sequence in base 2 indicates that its computation time is not truly linear with the sequence dimension, the additional one-time cost of determining the Gray bit has little impact on the near linear scaling observed in Figure 4.4. Most importantly the Niederreiter sequence in base 2 has the lowest computation cost of all the sequences tested because it has the lowest operation count. Specifically, the implementation of the Niederreiter sequence in base 2 in this investigation is 3 times faster than the random sequence for all sequence dimensions.

CHAPTER V

THE SIMULATION OF FREE MOLECULAR FLOW IN A TWO DIMENSIONAL DUCT

The intended goal of this investigation is to obtain an accurate and efficient simulation of low speed, non-equilibrium gas flows for microscale applications; and two approaches to achieve this goal are considered. The first approach is to apply empirical corrections to the Navier-Stokes equations in order to account for the non-equilibrium effects, which is covered in detail in Chapter II. The second approach is to develop a quasi-Monte Carlo (QMC) particle simulation, that achieves a faster error convergence rate than the $\mathcal{O}\left(N^{-1/2}\right)$ rate associated with DSMC (where N is the number of independent samples). In Chapter III, the theory behind general QMC integration is given, and the possible existence of a particle method with near linear error convergence $\mathcal{O}\left(N^{-1}(\log N)^{s-1}\right)$ is shown. QMC integration improves on the $\mathcal{O}\left(N^{-1/2}\right)$ convergence rate by replacing the random (or pseudo-random) sequence used in traditional Monte Carlo methods with a deterministic version, termed a low-discrepancy sequence, that attains a more uniform distribution throughout the integration domain. In Chapter IV, a new construction of the Weyl-Richtmyer sequence is presented, and the algorithmic implementation of all the low-discrepancy sequences tested in this investigation is reviewed as well. The purpose of Chapter V

is to develop the best possible QMC particle method for free molecular (collision-less) gas flow through a finite length duct in two dimensions. In particular, the principles used to develop the QMC particle simulation in this investigation are based on other successful QMC particle simulations constructed for model radiation transport and global luminosity problems. The resulting QMC particle method presented in this section for free molecular flow is then shown to have a superior error convergence rate than traditional DSMC.

The last 20 years has seen a surge in research for developing the method of QMC integration for several particle-type applications. Sarkar and Prasad [153] develop a QMC method for one dimensional particle transport through a solid medium with bi-directional scattering. Spanier [167] (with Li [168]) also develops QMC methods for model transport problems, and for a finite-state stochastic process represented by a Markov chain. QMC methods are constructed for ray-tracing applications by Keller [71, 72] for the problem of global luminosity, and by Kersch *et. al.* [74] for the radiative heat transfer found in semiconductor processing. Several stochastic systems in which the simulated particles follow a random-walk, or Brownian motion, have been solved using QMC methods. In particular, these QMC applications are developed by Caflisch and Moskowitz [23] for the Feynman-Kac integral, Morokoff and Caflisch [115] for one dimensional heat diffusion, and Moskowitz [119] for two quantum mechanical systems.[1]

Despite the potential for success illustrated in the aforementioned examples, it is the understanding of the author that no QMC particle method has ever demonstrated significant improvement over the DSMC simulation of the full Boltzmann equation, with respect to numerical convergence or computation time. Babovsky *et. al.* [8, 9]

[1] In [119], the quantum mechanical systems are simulated for the three dimensional harmonic oscillator, and the ground state energy of the helium atom.

propose a "low-discrepancy" selection procedure for choosing the particle collision pairs in a Nanbu-type[2] DSMC method. Unfortunately, the actual implementation of this low-discrepancy collision process into a full Boltzmann simulation creates a non-physical loss of particle energy in the simulations performed in [9].

Lécot and Coulibaly extend the concept of the low-discrepancy collision process from [8, 9], and apply it to several different model particle collision problems. In particular, the Boltzmann solution of the Krook and Wu problem[3] is studied by Lécot in [86, 87, 88]. Lécot is able to demonstrate that the low-discrepancy collision process simulated for the Krook and Wu problem does achieve a more uniform distribution of particle velocities than the Nanbu-type Monte Carlo simulation. The uniformity of the velocity distributions in [86, 87, 88] is defined in a specific sense using a general discrepancy measure; however, the connection of this discrepancy measure to the actual accuracy of the simulation for quantities of engineering interest (*e.g.* temperature) is not clear. Lécot and Coulibaly [89] more rigorously develop a QMC method based on a linearized collision model for a spatially homogeneous gas[4], which is demonstrated to have a slightly faster error convergence rate (*i.e.* $\mathcal{O}\left(N^{-0.55}\right)$ to $\mathcal{O}\left(N^{-0.60}\right)$) than DSMC. While the relative error of the QMC method is lower than traditional DSMC, the computational cost of the QMC method is also 3-6 times greater. Thus, it is not clear at which error levels the QMC method in [89] actually becomes faster than DSMC. Even though a QMC particle simulation of the Boltzmann equation with near-linear convergence is not demonstrated, the aforementioned research does establish many important mathematical proofs regarding

[2]See Nanbu's original paper [121] for the specific details on the method.

[3]Krook and Wu in [84] present an exact solution to the Boltzmann equation under the simplifying assumptions that the velocity distribution function is spatially uniform, and the collisions occur between Maxwell molecules.

[4]Coulibaly and Lécot also develop a QMC method for even simpler linear models of a Boltzmann-type equation in [33].

the theoretical convergence of a QMC simulation of the collision process.

The development of the QMC particle simulation in this investigation follows more of an engineering path than the previous work for the Boltzmann equation [8, 9, 33, 86, 87, 88, 89], which is more focused on the mathematical and theoretical aspects of the method. In particular, this investigation is concerned with the numerical convergence rate that can be achieved in practice by the QMC particle simulation; and the error level at which the QMC particle simulation is faster than traditional Monte Carlo. The goal of this chapter is to demonstrate if it is possible to develop a QMC particle simulation of free molecular (collision-less) duct flow with a near linear error convergence rate. The QMC Boltzmann simulations presented in [9, 33, 86, 87, 88, 89] are all for a spatially homogeneous, infinite expanse of gas. Thus, there is no simulation of the advection of particles or the stochastic particle-boundary interactions, which are both necessary for free molecular flow and the full Boltzmann equation. If it is not possible to achieve near-linear error convergence under the simplified condition of free molecular flow, it is unlikely that an efficient QMC particle simulation could be developed for the full Boltzmann equation.

Unlike the model collision simulations [9, 33, 86, 87, 88, 89], the QMC particle simulation of free molecular flow is actually very similar to the ray-tracing simulations used for global luminosity [71, 72], and radiative heat transfer [74]. In fact, the distribution of trajectory angles for the simulated particles is the same for all boundary reflections, regardless of the simulated particle representing a gas molecule, a light ray, or a packet of radiative energy. The QMC methods developed for the global luminosity and radiative heat transfer applications demonstrate a noticeable improvement in the error convergence over traditional Monte Carlo. Unfortunately, the fastest error convergence rate observed in [71, 72, 74] is only $\mathcal{O}\left(N^{-0.66}\right)$, which

is still less than the near-linear theoretical convergence. The results from these ray-tracing applications imply that it is most likely possible to construct an efficient QMC particle simulation for free molecular flow. However, the QMC method in [71, 72, 74] is only applied to a few specific ray-tracing problems. As a consequence, it is difficult to determine the possible magnitude of any performance gains achieved by QMC simulation for free molecular flow based solely on these other ray-tracing results.

There is also a significant difference between free molecular flow and the other ray-tracing applications, which further compounds the problem of making an accurate assessment about the QMC simulation of free molecular flow from the results in [71, 72, 74]. Unlike the ray-tracing simulations for global luminosity and radiative heat transfer, there is no absorption of the simulated particles at the wall surfaces in free molecular flow.[5] The presence of natural surface absorption is beneficial with respect to the performance of the QMC methods because it reduces the dimensionality of the problem. And as well-noted throughout the literature for many different applications [23, 74, 114, 116, 117, 118, 120, 146, 153, 167], the performance of the QMC method tends to decline as the problem dimension increases. Consequently, the QMC particle simulation of free molecular flow would likely have worse performance than observed in [71, 72, 74], if used for similar geometries. Because of the lack of conclusive evidence from the available literature, the purpose of this chapter is to demonstrate that it is indeed possible to create a QMC particle simulation for free molecular flow with near-linear error convergence.

In spite of achieving near-linear error convergence, the QMC particle simulation

[5]There are special cases of free molecular flow (*e.g.* chemical vapor deposition (CVD), and ionization) in which there is a natural absorption of the simulated particles; however, these cases are not considered in this investigation.

developed in this chapter is not the fastest simulation method for free molecular duct flow. While the QMC particle simulation is clearly faster than the traditional test particle Monte Carlo method, other techniques that directly approximate the probability distributions associated with the problem are the fastest. There may be specific flow problems (*e.g.* TMAC < 1) when the QMC particle is the fastest simulation technique because of the presence of singularities in the probability distribution. However, these particular cases are not the focus of this investigation. The overarching goal of this chapter and Chapter VI is to better understand how an efficient QMC particle simulation is constructed, not to develop the best overall method for free molecular duct flow. It is important to address the potential problems and limitations facing the QMC particle simulation for free molecular flow because they will be inherited in any simulation of the full Boltzmann equation. Therefore, the QMC particle simulation developed here is thoroughly tested in Chapter VI to determine its range of applicability, explore the dimension problems related to the method, and consider possible techniques to better avoid the limitations of the method.

An outline of the chapter organization is as follows. In Section 5.1, the fundamental probability distributions are derived for free molecular flow in a rectangular duct with a constant cross-section area. These probability distributions serve as the foundation on which all the simulation methods presented in this chapter are developed. The whole reason for developing a QMC particle simulation is motivated by the relatively slow convergence of the DSMC method. It is, therefore, very undesirable to use the DSMC solution to validate that the new QMC particle simulation has a faster error convergence rate and greater accuracy.[6] Instead, alternative meth-

[6]For the example presented in Section 5.5, the QMC particle simulation is able to achieve a relative error of 10^{-7} in less than 6 minutes on a 3.06 GHz Intel Xeon processor. In contrast, the traditional test particle Monte Carlo method would require over 10^{13} sample trajectories and at least 151 days to reach the same level accuracy on the same machine.

ods based on the simulation of the probability distributions directly are used for validating the performance gains of the QMC particle simulation. The construction details of these alternative simulations provide additional physical insight into the free molecular flow problem, which is useful in the development of the QMC particle simulation. As such, they are presented in Section 5.2 for the Markov chain simulation; Section 5.3 for the finite-state linear system simulation; and Section 5.4 for the Nyström method. Finally, the construction of the particle methods is reviewed in Section 5.5; and most importantly, the QMC particle simulation of free molecular duct flow is shown to achieve near-linear error convergence.

5.1 Basic Kinetics of Free Molecular Duct Flow

The study of internal[7] free molecular flows is one of the classic problems of gas kinetic theory. Estimation of the molecular flow rates through ducts and pipes was performed initially by Knudsen [76, 77], von Smoluchowski [181], and Dushman [43], both analytically and experimentally. The analysis of the molecular flow rates was extended further by Lorentz [102] and Clausing [30]. The results of Clausing are of specific interest to this investigation because they yield perhaps the best analytical estimate of the molecular flow rate through two dimensional ducts of finite length. A review of the early development of gas kinetic theory for these internal flows is given by Loeb in [98]. The first numerical simulations using the test-particle Monte Carlo method were performed by Davis [36] to calculate the free molecular flow rates through finite length pipes, concentric pipes, and elbows. A review of the performance and statistical properties of the test particle Monte Carlo method is

[7]The study of external free molecular flows is also a classic problem with many important applications to space system designs. A review of free molecular aerodynamics can be found for these applications in [54, 154].

given in the monograph of Bird [16].

The simulation techniques developed in this chapter for free molecular duct flow are based on the following problem assumptions. The duct inlet is assumed to be connected to a reservoir of infinite expanse, filled with a spatially homogeneous distribution of gas molecules in local thermodynamic equilibrium. Macroscopically, the temperature of the gas in the inlet reservoir is held constant, and the average gas velocity is zero. Furthermore, the gas density is assumed to be sufficiently low such that the mean free path of the gas molecules in the inlet reservoir is many times larger than the height of the duct. At the other end, the duct outlet is assumed to be connected to a reservoir of infinite expanse, which is held at a perfect vacuum. As a consequence of the low gas density throughout the system, any occurrence of inter-molecular collisions within the duct is exceedingly rare; and as such, they are neglected by the assumption of free molecular flow. The assumption that the reservoirs are of infinite expanse is important because it implies that these four additional conditions are also true: (i) the distribution of gas molecules entering the duct at the inlet plane is spatially uniform; (ii) the flow conditions inside the two reservoirs are constant in time; (iii) the gas molecules that escape the duct have no local effect on any new molecules that may enter the duct interior; and (iv) the gas molecules that escape the duct have no global effect on the flow conditions inside the reservoirs themselves.

In addition to these assumptions about the flow conditions in the problem, there are also the following assumptions based on the duct geometry. Specifically, the duct has a finite length ℓ and a constant rectangular cross-section with a height h and width w. The duct width is many times larger than the height ($w \gg h$); thus, any flow changes in the direction of the width are negligible, and the geometry

can be considered two dimensional. The surfaces of the duct wall are sufficiently rough such that any gas molecule colliding with the wall is assumed to undergo a diffuse reflection with the surface. During a collision, a gas molecule is assumed to remain in contact with the surface for a long enough time as to fully accommodate to the thermodynamic conditions at the wall. Stated alternatively, the wall surfaces of the duct are assumed to have a tangential momentum accommodation coefficient (TMAC) and a thermal accommodation coefficient equal to one (see [109, 179, 54, 154] and Chapter II). Finally, the temperature of the duct walls is assumed to be constant everywhere and equal to the inlet reservoir temperature. Thus, any local heating or cooling at the surface caused by the energy exchange between the gas molecules and the wall is neglected.

There is only one flow quantity of interest that is calculated for the simulation methods developed in this chapter for internal free molecular flow. This quantity, which is denoted by Ψ, is the probability a particle enters the duct at the inlet and eventually escapes the duct through the outlet after any number (possibly infinite) of collisions with the interior walls of the duct. Given a fixed interval of time Δt, let N_{tot} denote the total number of particles that enter the duct from the inlet reservoir in time Δt. The total number of particles N_{tot} entering the duct can be subdivided into three categories: (i) N_{in} is the number of particles that escape the duct through the inlet in time Δt; (ii) N_{out} is the number of particles that escape the duct through the outlet in time Δt; and (iii) N_{duct} is the number of particles that have yet to escape and still remain within the duct after time Δt. Once a specific particle escapes the duct interior it is not allowed to re-enter the domain; hence, $N_{tot} = N_{in} + N_{out} + N_{duct}$. The free molecular flow quantity Ψ can then be considered, in a more physical context, as the fraction of the total particles entering the duct that eventually escape the duct

through the outlet. That is,

$$\Psi = \lim_{\Delta t \to \infty} \frac{N_{out}}{N_{tot}}.$$

Once the non-dimensional flow quantity Ψ is known for a particular duct geometry, it can be used to calculate a wide range of dimensional mass flow rates for the geometry. This is assuming, of course, that the gas density at the inlet and outlet reservoirs is sufficiently low enough relative to the length scales of the duct for the gas flow to be considered in the free molecular regime.[8]

There does not seem to be a consensus in the literature over the exact name for the free molecular flow quantity Ψ. In [43], Dushman borrows from the electrical analogue of the problem and refers to its dimensional form as the "conductivity," which measures the ease at which particles are able to flow through the duct geometry. In [30], Clausing defines Ψ as the *Durchlaufswahrscheinlichkeit* which loosely translates into the "probability of running through [the duct]." Since both of these names lend some insight into the physical nature of the flow quantity being simulated, as a compromise, Ψ is referred to as the "conductance probability" throughout this investigation. Actual calculations of the conductance probability Ψ have been performed by Clausing [30], using an approximate analytical solution for finite length ducts and pipes; and by Davis [36], using the test particle Monte Carlo simulation for cylindrical pipes and elbows.

The collision-less Boltzmann equation[9] provides the mathematical description for free molecular gas flow (see [16, 54, 154, 179] for more examples). Let the distribution function $F = F(\mathbf{x}, \mathbf{v}, t)$ represent the number of particles located within the

[8]Note that Ψ can also be used to calculate the net mass flow rate when both the inlet and outlet reservoirs contain some finite gas density because the particle-particle interactions are ignored in the free molecular regime. Thus, in this regime, the flow from the outlet to the inlet is independent of the flow from the inlet to the outlet, and both can be calculated from Ψ.

[9]Also referred to as the Vlasov equation.

infinitesimal neighborhood of the spatial location $\mathbf{x} = (x_1, x_2, x_3)$ that travel at a velocity in the infinitesimal neighborhood of $\mathbf{v} = (v_1, v_2, v_3)$ at time t. The evolution of this distribution function $F(\mathbf{x}, \mathbf{v}, t)$ is governed by the collision-less Boltzmann equation. Assuming that there are no external forces present in the free molecular flow, the collision-less Boltzmann equation is given by

$$\frac{\partial F}{\partial t} + v_1 \frac{\partial F}{\partial x_1} + v_2 \frac{\partial F}{\partial x_2} + v_3 \frac{\partial F}{\partial x_3} = 0. \tag{5.1}$$

However, the calculation of the conductance probability Ψ, based on the previously stated assumptions, does not actually require the collision-less Boltzmann equation (5.1) to be solved formally.[10] In fact, all the conductance probability simulations developed in this chapter are constructed solely from the physical behavior of the gas molecules at the boundaries of the duct, without ever using the collision-less Boltzmann equation in (5.1) directly. This should not be too surprising because the collision-less Boltzmann equation is, after all, a linear hyperbolic equation with a standard method of characteristics type solution. The information traveling along the solution characteristics, in this case, relates to the number of gas molecules in a particular region of velocity space. In addition, these solution characteristics are independent of each other because no inter-molecular collisions are assumed to occur in free molecular flow. Therefore, the information traveling along the characteristics remains constant, and is determined entirely by the boundary conditions of the problem.

There are three types of boundary conditions for free molecular duct flow: (i) the inflow boundary condition; (ii) the outflow boundary condition; and (iii) the diffuse gas molecule reflections at the duct wall. The inflow boundary condition is

[10] A more formal treatment of the solutions of the Boltzmann equation can be found in the excellent monographs of Cercignani [25] and Kogan [80]

determined by the velocity distribution of the gas molecules crossing the plane of the duct from the inlet reservoir. Because of the assumption of local thermodynamic equilibrium and zero bulk velocity, the velocity distribution of the gas molecules in the inlet reservoir follows a Maxwellian distribution $\mathcal{M}(\mathbf{v})$. That is,

$$\mathcal{M}(\mathbf{v}) = \left(\frac{m}{2\pi kT}\right)^{3/2} \exp\left(-\frac{m|\mathbf{v}|^2}{2kT}\right) \quad \text{for } v_1, v_2, v_3 \in (-\infty, \infty), \tag{5.2}$$

where $\mathbf{v} = (v_1, v_2, v_3)$ is the velocity of the gas molecule, m is the mass of a single gas molecule, k is Boltzmann's constant, and T is the temperature in the inlet reservoir. Let $\Phi_p(\mathbf{v})$ denote the velocity distribution of the gas molecules crossing the plane of the duct from the inlet reservoir, which is then given by

$$\Phi_p(\mathbf{v}) = 2\left(\frac{\pi m}{2kT}\right)^{1/2} v_1 \mathcal{M}(\mathbf{v}) \text{ for } v_1 \in (0, \infty) \text{ and } v_2, v_3 \in (-\infty, \infty), \tag{5.3}$$

where v_1 is the velocity component normal to the inlet plane (positive direction pointing toward the duct). Note the additional velocity term in (5.3) is to account for the probability of a gas molecule actually fluxing across the inlet plane from the gas reservoir in an infinitesimal span of time. The probability distribution $\Phi_p(\mathbf{v})$ is constant everywhere on the inlet plane, and the flux of particles from the duct to the reservoir across the inlet plane has no effect on the boundary condition, based on the previous assumptions about the inlet reservoir. Thus, the velocity distribution $\Phi_p(\mathbf{v})$ (5.3) of the gas molecules crossing the plane of the duct from the inlet reservoir completely defines the necessary inflow boundary condition to determine the conductance probability Ψ of the duct.[11]

The outflow boundary condition is located at the plane connecting the duct to the outlet reservoir, which makes this boundary condition the simplest because the

[11] If the dimensional mass flow rate is to be calculated for the duct, then the particle number density inside the inlet reservoir is necessary as well.

reservoir is a perfect vacuum. As a consequence, there is zero flux of particles across the outlet plane from the vacuum to the duct. Moreover, the flux of particles from the duct to the reservoir across the outlet plane has no effect on the outflow boundary condition, based on the previously stated assumptions about the reservoirs.

The third boundary condition, the gas molecule reflections at the duct wall, is often evaluated in terms of the following model description of the physical collision process for a fully diffuse wall. When a gas molecule first collides with a diffuse wall, it is assumed to be temporarily absorbed by molecular structure of the wall surface. During this absorption period, the energy of the gas molecule relaxes, or accommodates, to the same thermodynamic state as the wall surface. After which point the gas molecule is re-emitted from the wall surface into the duct as if it is crossing the wall plane from a reservoir of gas molecules in local thermodynamic equilibrium with the wall temperature. Classic results from molecular beam experiments indicate that there is actually a finite absorption period.[12] However, in this investigation, the absorption, accommodation, and re-emission processes of the colliding gas molecules are assumed to occur instantaneously. Given a fixed interval of time Δt, the number of molecules re-emitted at a particular wall region in Δt is equal to number of molecules colliding with the same region in Δt. Since the wall temperature is the same as the inlet reservoir temperature, the velocity distribution of the re-emitted gas molecules crossing the wall plane is the same as the inflow boundary condition $\Phi_p(\mathbf{v})$ (5.3).

Now that the boundary conditions of the free molecular duct flow are defined, it is possible to further simplify them for the specific calculation of the conductance probability Ψ. In particular, the duct geometry is assumed to be two dimensional,

[12]For example, an average absorption time of $3 \cdot 10^{-5}$ seconds is found for argon gas in a glass capillary tube at 90°K (see the review of molecular beam given by de Boer in [38]).

which can be used to simplify the inflow and the diffuse wall reflection boundary condition $\Phi_p(\mathbf{v})$ in (5.3). If the v_3 velocity component is taken to be in the direction of the duct width, the probability distribution function $\Phi_p(\mathbf{v})$ can be integrated over all possible velocities, $v_3 \in (-\infty, \infty)$, to yield a distribution function for the two critical dimensions of the duct geometry. Specifically,

$$\Phi_p(v_1, v_2) = \frac{2}{\sqrt{\pi}} \left(\frac{m}{2kT}\right)^{3/2} v_1 \exp\left(-\frac{m}{2kT}(v_1^2 + v_2^2)\right),$$

where $v_1 \in (0, \infty)$ in the direction normal to the plane, and $v_2 \in (-\infty, \infty)$ in the direction parallel to the plane. The distribution Φ_p can be further simplified using a polar transformation of the Cartesian velocity components (v_1, v_2) into a two dimensional speed v_r and a trajectory angle θ (measured from the normal of the plane). After the coordinate transformation, the velocity distribution of the gas molecules crossing the plane into the duct (either from the inlet reservoir or a wall reflection) is given by

$$\Phi_p(v_r, \theta) = \frac{2}{\sqrt{\pi}} \left(\frac{m}{2kT}\right)^{3/2} v_r^2 \cos\theta \exp\left(-\frac{m}{2kT}(v_r^2)\right), \tag{5.4}$$

where $v_r \in (0, \infty)$ and $\theta \in (-\frac{\pi}{2}, \frac{\pi}{2})$. Note that an additional velocity term v_r is included in (5.4) to account for the Jacobian of the coordinate transformation for the differential elements of velocity space $dv_1 dv_2$ and $v_r dv_r d\theta$ associated with their respective distribution functions.

The velocity distribution (5.4) of the gas molecules crossing the plane into the duct can be separated into two independent functions $f_{v_r}(v_r)$ and $f_\theta(\theta)$ such that

$$\Phi_p(v_r, \theta) = f_{v_r}(v_r) f_\theta(\theta).$$

The trajectory angle θ of a gas molecule entering the duct from the inlet reservoir (or reflecting from the duct wall) is therefore independent of its two dimensional

speed v_r. The velocity (v_r, θ) of a gas molecule is not affected by body forces and collisions while traveling within the duct interior, as indicated by the collision-less Boltzmann equation (5.1). Thus, the entire trajectory is determined by the points where the gas molecule intersects with the boundaries (inflow, outflow, and wall reflections) of the duct. These points along the boundary are found solely from the trajectory angle and the specific duct geometry, and are independent of the two dimensional molecular speed v_r. In addition, the calculation of the conductance probability Ψ only depends on the final location of the gas molecule trajectory (either an intersection with the inlet plane or the outlet plane). Therefore, the calculation of the conductance probability Ψ under the previously stated assumptions for the free molecular duct flow is only dependent on the distribution of trajectory angles $f_\theta(\theta)$ at the boundary planes. By integrating the molecular speed v_r over all possible values $(0, \infty)$ in (5.4), the distribution of trajectory angles $f_\theta(\theta)$ is given by,

$$f_\theta(\theta) = \frac{1}{2}\cos\theta \quad \text{for } \theta \in (-\tfrac{\pi}{2}, \tfrac{\pi}{2}). \tag{5.5}$$

Note that the angle θ is measured relative to the normal of the boundary plane.

From the standpoint of a Monte Carlo particle simulation, it is very easy to generate sample trajectory angles from the cosine distribution in (5.5). However, for the two dimensional duct geometry, it is far more efficient in terms of computation time to work with the particle position on the boundaries of the duct instead of the actual trajectory angle. By avoiding the direct usage of the particle trajectory angle, the resulting Monte Carlo simulation is able to eliminate the costly evaluations of trigonometric and inverse trigonometric functions in favor of simple square root calculations. Furthermore, it is easier to develop all the simulation methods considered in this investigation for free molecular duct flow, including the non-particle based

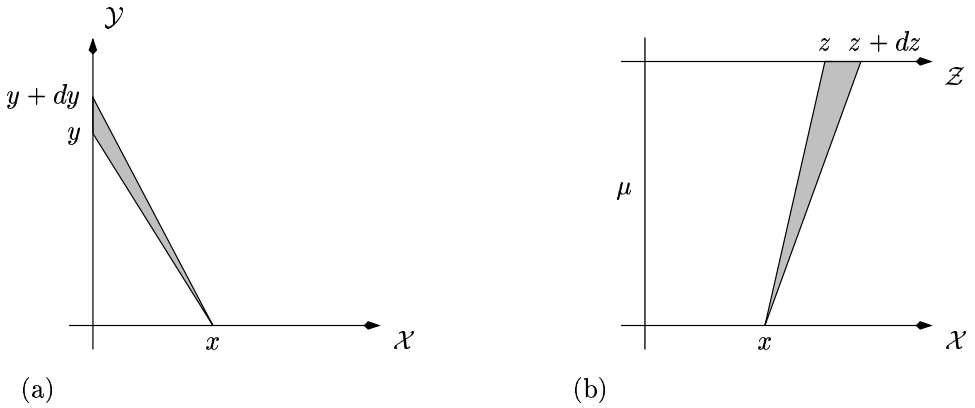

Figure 5.1: Illustration of the two basic probability distributions used in the simulation of free molecular duct flow: (a) the perpendicular transition probability $\mathcal{T}_\perp(x, y)$; and (b) the parallel transition probability $\mathcal{T}_\parallel(x, z; \mu)$.

techniques, using the particle position on the domain boundary instead of the trajectory angle. In order to simulate free molecular duct flow in two dimensions, two types of particle moves from the domain boundary must be considered: (i) moves between the perpendicular boundaries; and (ii) moves between the parallel boundaries. These moves are direct, *i.e.* only the first boundary intersected by a particle is considered. Multiple reflections between boundaries must be treated as a series of single moves by the particle from boundary to boundary.

There are two basic probability distributions corresponding to the two types of particle moves from the domain boundary. The first probability distribution is the *perpendicular transition probability*, which is denoted by $\mathcal{T}_\perp(x, y)$; an illustration of $\mathcal{T}_\perp(x, y)$ is given in Figure 5.1(a). Given two perpendicular lines $\mathcal{X}$ and $\mathcal{Y}$, let x denote the distance of a particle on line $\mathcal{X}$ from the intersection $\mathcal{X} \cap \mathcal{Y}$. The perpendicular transition probability is defined such that $\mathcal{T}_\perp(x, y)dy$ is the probability a particle at $x \in \mathcal{X}$, with a trajectory angle following a cosine distribution, directly intersects the line $\mathcal{Y}$ in the infinitesimal interval $[y, y + dy)$. Here y is measured from the same intersection $\mathcal{X} \cap \mathcal{Y}$ as x. The perpendicular transition probability $\mathcal{T}_\perp(x, y)$

is then calculated by integrating the cosine distribution of the trajectory angle over the appropriate interval, that is

$$\lim_{dy\to 0} \mathcal{T}_{\perp}(x,y)dy = \lim_{dy\to 0} \int_{\tan^{-1}\left(\frac{x}{y+dy}\right)}^{\tan^{-1}\left(\frac{x}{y}\right)} \frac{1}{2}\cos\theta d\theta.$$

Taking the appropriate limits, one obtains

$$\mathcal{T}_{\perp}(x,y) = \frac{xy}{2(x^2+y^2)^{3/2}}. \tag{5.6}$$

It is important to note that the symmetry of the trajectory angle distribution $f_{\theta}(\theta)$ in (5.5) implies that the result for $\mathcal{T}_{\perp}(x,y)$ in (5.6) is valid for particles on both sides of the line $\mathcal{Y}$, provided the distance x is measured in a positive sense for both cases.

Similar to $\mathcal{T}_{\perp}(x,y)$, the second probability distribution used in the construction of the simulation methods is the *parallel transition probability*, which is denoted by $\mathcal{T}_{\parallel}(x,z;\mu)$; an illustration of $\mathcal{T}_{\parallel}(x,z;\mu)$ is given in Figure 5.1(b). Given two parallel lines $\mathcal{X}$ and $\mathcal{Z}$ that are separated by a distance ℓ, let x denote the distance of a particle on line $\mathcal{X}$ from some reference line perpendicular to both $\mathcal{X}$ and $\mathcal{Z}$. The parallel transition probability is defined such that $\mathcal{T}_{\parallel}(x,z;\mu)dz$ is the probability a particle at $x \in \mathcal{X}$, with a trajectory angle following a cosine distribution, directly intersects the line $\mathcal{Z}$ in the infinitesimal interval $[z, z+dz)$. Here z is the distance relative to the same perpendicular line as x. The parallel transition probability $\mathcal{T}_{\parallel}(x,z;\mu)$ is then calculated by integrating the cosine distribution of the trajectory angle over the appropriate interval, that is

$$\lim_{dz\to 0} \mathcal{T}_{\parallel}(x,z;\mu)dy = \lim_{dz\to 0} \int_{\tan^{-1}\left(\frac{x-z-dz}{\mu}\right)}^{\tan^{-1}\left(\frac{x-z}{\mu}\right)} \frac{1}{2}\cos\theta d\theta.$$

Taking the appropriate limits, one obtains

$$\mathcal{T}_{\parallel}(x,z;\mu) = \frac{\mu^2}{2\big((x-z)^2+\mu^2\big)^{3/2}}. \tag{5.7}$$

Note that the parallel transition probability only depends on the magnitude of $|x-z|$ and not the direction because of the symmetry of the trajectory angle distribution $f_\theta(\theta)$ in (5.5).

5.2 Markov Chain Simulation

The Markov chain simulation is perhaps the most straightforward, non-particle method for calculating the conductance probability. A Markov chain represents the stochastic behavior of a given population moving at random between a finite number of states. While the movement between states is not deterministic, the probability $p \in [0, 1]$ is known for the transition of a population member from one state to another. Since the Markov chain represents a stochastic process satisfying the Markov property (see [47] for more details), the evolution of the population is discrete in the following sense. Given an initial realization of the population P_0 (*i.e.* the initial distribution of population members between the system states), any future realization P_n depends only on the realization that immediately precedes it P_{n-1} for all $n \geq 1$. Furthermore, for the simulation considered here, the transition probability between states is constant throughout the evolution of the population; and as such, the underlying stochastic process is referred to as a discrete-time, time-homogeneous Markov chain.[13]

In order to simulate the conductance probability as a Markov chain, the population under consideration is taken to be the gas molecules that enter the duct from the inlet. The boundaries associated with the duct, the inlet plane, outlet plane, and interior walls, are taken to be the finite states of the system. A gas molecule belongs

[13]The reference to time is an artifact of its common usage for stochastic processes evolving in time. However, the terminology is still used for any ordered set of realizations of the population $(P_0, P_1, P_2, \ldots)$, that satisfies the aforementioned conditions; regardless of the actual physical processes responsible for changes in the population.

to the state corresponding to its last intersection with the boundary of the duct. The duct inlet plane and outlet plane are each treated as single states, while the interior duct wall is divided into N distinct intervals of uniform size that cover the entire wall boundary. As the number of interior states N increases, so too does the accuracy of the Markov Chain simulation for calculating the conductance probability. The transition probability between two states is simply the probability a gas molecule leaving the boundary region of the first state will next intersect the boundary at a region corresponding to the second state. The evolution of the Markov chain is not consistent with physical time because the gas molecules, in general, have different transit times between boundary intersections. However, the gas flow is collision-less which means there is no interaction between molecules during the transition time between boundary interaction. Hence, the Markov chain is able to decouple the stochastic evolution of the gas molecules from their individual transit times while still remaining physically accurate. For the simulation of the conductance probability, one is interested in solving for the expected (or average) long-term behavior of the stochastic system. Specifically, the conductance probability is determined from the long-term average fraction of the gas molecules that enter the duct through the inlet plane that then eventually escape through the outlet plane.

The expected behavior of any Markov chain can be represented by a linear system,

$$\mathbf{v}^{(n)} = A\mathbf{v}^{(n-1)}, \tag{5.8}$$

where the vector $\mathbf{v}^{(n)}$ represents the expected distribution of the population among the states at the n^{th} realization of the system, and the matrix A (also known as the Markov matrix), represents the transition probabilities between the system states. For the Markov chain simulation of the conductance probability with N interior wall

states, each element $v_i^{(n)}$ of the vector $\mathbf{v}^{(n)} \in \mathbb{R}^{N+2}$ in (5.8) represents the average number of gas molecules occupying state i at the n^{th} realization of the system. Similarly, each element A_{ij} of the matrix $A \in [0,1]^{(N+2)\times(N+2)}$ in (5.8) represents the probability that a molecule leaving state j will undergo its next intersection with the duct boundaries within state i. Each realization n of the system corresponds to the distribution of gas molecules among the states of the duct after undergoing n intersections with the boundary (or equivalently, n particle moves).

Since the conductance probability Ψ only involves the fraction of molecules that eventually escape the duct through the outlet, the vector $\mathbf{v}^{(n)}$ need not represent the total number of gas molecules in each state. Instead, it is sufficient to only consider the average fraction of the total number of gas molecules. That is, the vector $\mathbf{v}^{(n)}$ is normalized such that $v_i^{(n)}$ now represents the probability a gas molecule is in state i after n moves. It is this representation of the $\mathbf{v}^{(n)} \in [0,1]^{N+2}$, as the average fraction of molecules distributed among the system states, that is assumed for the rest of the investigation. Note that all the matrix and vectors in (5.8) now represent actual probabilities; and thus, their components are restricted to real numbers in the unit interval $[0,1]$. Given an initial probability distribution $\mathbf{v}^{(0)}$ of the gas molecules entering from the inlet, the expected distribution of the molecules among the states of the duct after n moves is determined by simply repeating the matrix-vector multiplication in (5.8). The conductance probability is therefore found by calculating long-term behavior of the stochastic system (*i.e.* $\lim_{n\to\infty} \mathbf{v}^{(n)}$).

The specific construction of the transition matrix A and initial probability distribution vector $\mathbf{v}^{(0)}$ are determined from the perpendicular $\mathcal{T}_\perp$ and parallel $\mathcal{T}_\parallel$ transition probabilities derived in Section 5.1. The transition matrix A is divided into several different blocks, each corresponding to different physical processes within

the stochastic system. This organization serves to facilitate the development of the Markov chain simulation, and to establish a consistent notation between the other simulation techniques presented later in this chapter. Assuming States $1, \ldots, N$ represent the interior states of the duct wall, the matrix A is defined by the following,

$$A = \begin{pmatrix} K_{11} & \cdots & K_{1N} & 0 & 0 \\ \vdots & & \vdots & \vdots & \vdots \\ K_{N1} & \cdots & K_{NN} & 0 & 0 \\ g_1 & \cdots & g_N & 1 & 0 \\ h_1 & \cdots & h_N & 0 & 1 \end{pmatrix}, \tag{5.9}$$

where the components K_{ij} form an $N \times N$ matrix K that represents the transition probability between the interior wall states only, the components g_i form a vector $\mathbf{g}$ that represents the probability a gas molecule at the wall state i intersects the inlet boundary (State N+1) during its next move, and the components h_i form a vector $\mathbf{h}$ that represents the probability a gas molecule at the wall state i intersects the outlet boundary (State N+2) during its next move. An illustration of the different components of the transition matrix A in (5.9) is given in Figure 5.2.

There are two important points to be noted about the transition matrix A that are true in general for any Markov matrix. First, there is no loss of gas molecules from the system. That is, if a gas molecule begins the simulation in some initial state prescribed by $\mathbf{v}^{(0)}$, then it must remain in one of the system states for all subsequent realizations of the population $\mathbf{v}^{(n)}$ for $n \geq 1$. Since the initial distribution $\mathbf{v}^{(0)}$ is normalized to represent the average fraction of particles in each state (*i.e.* $\sum_{i=1}^{N+2} v_i^{(0)} = 1$), then all future realizations of $\mathbf{v}^{(n)}$ for $n = 1, 2, \ldots$ also satisfy the normalization condition. The existence of this condition is confirmed by the fact that the sum of every column of matrix A (5.9) is equal to one. Second, the original

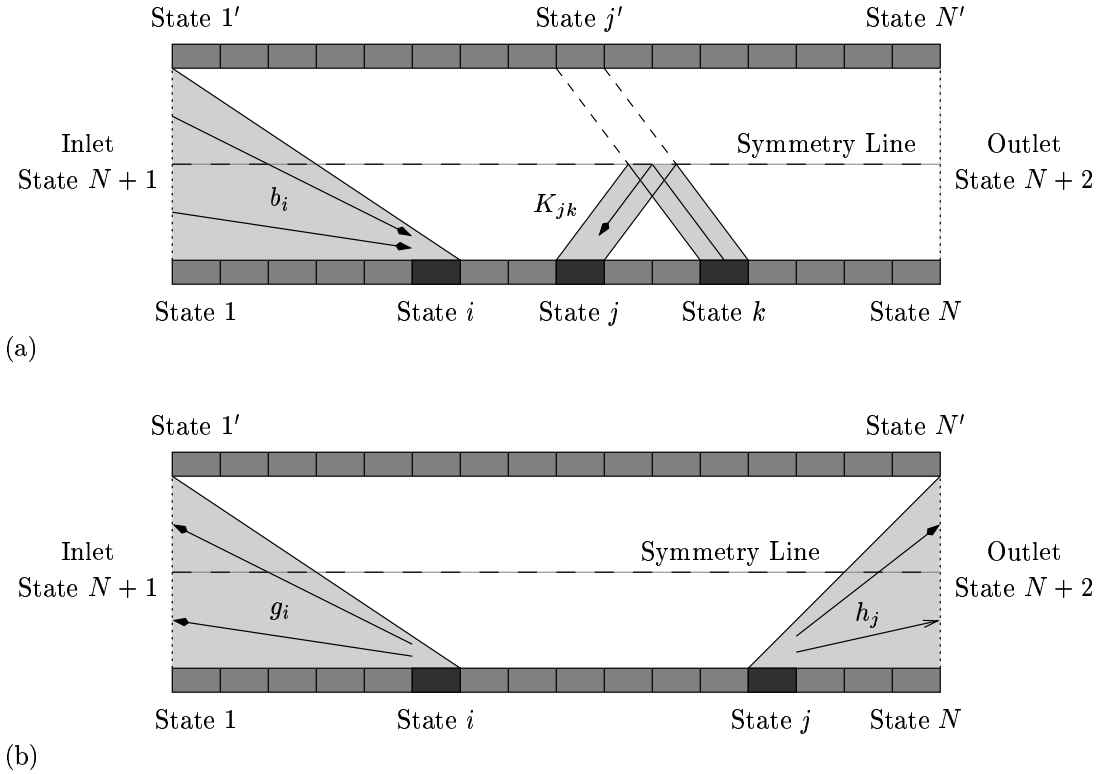

Figure 5.2: Illustration of the different types of transition probabilities for the gas molecules in the Markov chain simulation: (a) the initial probability distribution of molecules from the inlet **b**, and the transition probability matrix K between the interior wall states of the duct; and (b) the probability of escaping the interior wall states through the inlet **g** and outlet **h** states.

assumptions for the free molecular duct flow problem eliminate the possibility of the gas molecules re-entering the duct once they have escaped through either the inlet or outlet planes. Thus, when a gas molecule enters either the inlet or outlet state it is essentially trapped in that state, unable to leave for the rest of the simulation. In general, a state within the Markov chain that does not allow its members to leave is referred to as an absorbing state. An absorbing state appears in the transition matrix A in (5.9) as a column of zeros except for a one on the matrix diagonal.

In a manner similar to the transition matrix A in (5.9), the initial probability distribution $\mathbf{v}^{(0)}$ is also divided into two parts corresponding to different stochastic processes. The initial probability distribution $\mathbf{v}^{(0)}$ is given by

$$\mathbf{v}^{(0)} = (b_1, \ldots, b_N, 0, \rho(L))^T, \tag{5.10}$$

where the elements b_i form a vector $\mathbf{b}$ that represents the probability of a gas molecule entering the duct through the inlet and first striking the wall at the i^{th} interior state, and $\rho(L)$ represents the fraction of particles that reach the outlet directly from the inlet without ever colliding with the walls of the duct. An illustration of the initial probability distribution of molecules from the inlet $\mathbf{b}$ is given in Figure 5.2.

Before proceeding with the actual calculation of the transition matrix A and the initial probability distribution $\mathbf{v}^{(0)}$, it is necessary to first address the dimensional scale and symmetry of the duct geometry. The calculation of the conductance probability only depends on the distribution of trajectory angles $f_\theta(\theta)$ (5.5) for the gas molecules leaving the boundaries of the duct, based on the development in Section 5.1. The actual dimensional scale of the duct geometry is not needed to calculate the next intersection of a gas molecule with the duct boundary, when only using the trajectory angle θ of the gas molecules. Instead, the ratio of the duct length

ℓ and duct height h is all that is necessary to completely determine the boundary intersection points of the gas molecule from the trajectory angle. Therefore, the dimensions of all the duct geometries considered in this investigation are normalized by the duct height. This implies that all geometries have a height equal to one, and length $L = \ell/h$ equal to the duct length to height ratio.

Geometrically, there are two planes of symmetry in the rectangular ducts illustrated in Figure 5.2 for the free molecular flow: (i) a left-right symmetry plane between the inlet and outlet planes; and (ii) a top-bottom symmetry plane between the upper and lower duct walls. It is not possible to exploit the left-right symmetry plane in the calculation of the conductance probability. While the inlet and outlet states are symmetric in a geometric sense, the physical processes occurring at the two states is vastly different. As such, both the inlet and outlet must be included in any valid calculation of the conductance probability. The top-bottom symmetry plane, however, can be exploited in the calculation of the conductance probability. This is due to the fact that the distribution of trajectory angles $f_\theta(\theta)$ (5.5) is symmetric about the surface normal, which makes the walls physically indistinguishable from each other.[14] It is wasteful, in terms of computation time and memory, to include both the upper and lower walls in the Markov chain simulation; in fact, this is true for all the simulations developed in this chapter. Therefore, only one duct wall and the top-bottom symmetry plane are simulated during the calculation of the conductance probability.

All the quantities needed to construct the transition matrix A in (5.9) and the initial probability distribution $\mathbf{v}^{(0)}$ in (5.10) for the Markov chain simulation are

[14]The upper and lower duct walls are indistinguishable in terms of both the initial probability distribution of gas molecules first intersecting the duct wall from the inlet, and the diffuse reflection of gas molecules from the wall surface.

determined from the basic transition probabilities developed in Section 5.1. More specifically, the perpendicular transition probability $\mathcal{T}_\perp(x,y)$ in (5.6), and the parallel transition probability $\mathcal{T}_\parallel(x,z;\mu)$ in (5.7) are integrated over the appropriate intervals to obtain the necessary probabilities for A and $\mathbf{v}^{(0)}$. Let $0 = x_1 < x_2 < \cdots < x_{N+1} = L$ denote the endpoints of the interior wall states such that the interval $[x_i, x_{i+1})$ is the location of the i^{th} state. Note that all the interior wall states are the same length $\Delta x = x_{i+1} - x_i = L/N$, for $1 \leq i \leq N$. The probability K_{ij} of a gas molecule leaving the wall at state j and intersecting the wall at state i is found from the parallel transition probability $\mathcal{T}_\parallel(x,z;\mu)$, with the separation distance $\mu = 1$ set to the non-dimensional duct height. That is,

$$\begin{aligned} K_{ij} &= \frac{1}{\Delta x}\int_{x_j}^{x_{j+1}}\int_{x_i}^{x_{i+1}} \mathcal{T}_\parallel(x,z;1)\,dz\,dx \\ &= \frac{1}{2\Delta x}\Big(\sqrt{1+(x_{i+1}-x_j)^2} - \sqrt{1+(x_i-x_j)^2} \\ &\qquad -\sqrt{1+(x_{i+1}-x_{j+1})^2} + \sqrt{1+(x_i-x_{j+1})^2}\Big). \end{aligned} \tag{5.11}$$

Note that the function $\mathcal{T}_\parallel(x,z;\ell)$ is symmetric in its arguments, *i.e.* $\mathcal{T}_\parallel(x,z;\ell) = \mathcal{T}_\parallel(z,x;\ell)$. Thus, the integration order in (5.11) can also switched to obtain $K_{ij} = K_{ji}$, which implies that the matrix K is symmetric. The probability of reaching state i from state j from a reflection about the symmetry line is the same as reaching the mirror image of state i on the opposing wall, as illustrated in Figure 5.2(a) (note that the non-simulated, or ghost states, are denoted with a prime). Therefore, the calculation of the interior state transition probability K is the same for the one-wall and two-wall Markov chain simulations.

The remaining quantities needed to construct the transition matrix A are found in the same manner as K. The probability g_i of a gas molecule escaping the wall at state i through the inlet is found from the perpendicular transition probability

$\mathcal{T}_\perp(x,y)$ from (5.6),

$$\begin{aligned} g_i &= \frac{1}{\Delta x}\int_{x_i}^{x_{i+1}}\int_0^1 \mathcal{T}_\perp(x,y)dydx \\ &= \frac{1}{2} - \frac{1}{2\Delta x}\left(\sqrt{x_{i+1}^2+1} - \sqrt{x_i^2+1}\right). \end{aligned} \tag{5.12}$$

Because the distribution of trajectory angles f_θ in (5.5) is symmetric, the probability h_i of a gas molecule escaping the wall at state i through the outlet is the same as the calculation for g_i except with the integration variable x in (5.12) replaced by $L-x$. Hence,

$$\begin{aligned} h_i &= \frac{1}{\Delta x}\int_{x_i}^{x_{i+1}}\int_0^1 \mathcal{T}_\perp(L-x,y)dydx \\ &= \frac{1}{2} - \frac{1}{2\Delta x}\left(\sqrt{(x_i-L)^2+1} - \sqrt{(x_{i+1}-L)^2+1}\right). \end{aligned} \tag{5.13}$$

Note that as a result of the problem symmetry $h_i = g_{N+1-i}$, for $1 \le i \le N$. For the two-wall Markov chain simulation, the escape probability (inlet or outlet) must be considered for both walls, which would double the total escape probability for the one-wall Markov chain simulation. However, the number of gas molecules in each state of the two-wall simulation is half of the one-wall simulation. Thus, the calculation of g_i and h_i is the same for both the one-wall and two-wall Markov chain simulations.

Continuing in the same manner, the quantities needed for the initial probability distribution $\mathbf{v}^{(0)}$ in (5.10) are also calculated. The probability b_i of a gas molecule entering the duct through the inlet and directly intersecting the wall at state i is found from the perpendicular transition probability,

$$\begin{aligned} b_i &= 2\int_0^1\int_{x_i}^{x_{i+1}} \mathcal{T}_\perp(x,y)dydx \\ &= x_{i+1} - x_i + \sqrt{x_{i+1}^2+1} - \sqrt{x_i^2+1}. \end{aligned} \tag{5.14}$$

The initial probability distribution for the interior states is the only quantity in A and $\mathbf{v}^{(0)}$ that is different between the one-wall and two-wall Markov chain simulations. For the one-wall simulation, all the molecules that would normally reach state i and state i' in the two-wall simulation are combined into a single state, which is why the factor of 2 appears in front of the integral in (5.14). Thus, the value of b_i for the one-wall simulation is exactly twice that of b_i for the two-wall simulation, for the same state (either i or i').

The final quantity needed for the Markov chain simulation is the probability $\rho(L)$ that a molecule entering the duct at the inlet will directly escape through the outlet without ever colliding with the wall. The direct escape probability $\rho(L)$ is found from the parallel transition probability $\mathcal{T}_{\parallel}(x, z; \mu)$, with the separation distance $\mu = L$ set to the non dimensional duct length to height ratio,

$$\begin{aligned} \rho(L) &= \int_0^1 \int_0^1 \mathcal{T}_{\parallel}(x, z; L) dz dx \\ &= \sqrt{L^2 + 1} - L. \end{aligned} \tag{5.15}$$

Note that the probability a molecule escapes the duct without colliding with the wall tends to zero as the duct length increases, which is consistent with the expected behavior of free molecular gas flow.

At this point all the matrix and vector quantities in (5.8) are defined, and now the focus can shift to the calculation of the long-term behavior of the Markov chain for the simulation of the conductance probability. Let Ψ_n denote the probability that a particle escapes the duct through the outlet plane while only colliding with the interior walls n or fewer times. In this investigation (see Figure 5.2), the outlet state corresponds to the $(N+2)^{th}$ coordinate $v_{N+2}^{(n)}$ of the gas molecule probability

distribution $\mathbf{v}^{(n)}$. Hence,

$$\Psi_n = v_{N+2}^{(n)} = \mathbf{e}_{N+2}^T A^n \mathbf{v}^{(0)}, \tag{5.16}$$

where $\mathbf{e}_i$ is the elementary vector consisting of all zeros except for the i^{th} component which is one. As noted earlier, the conductance probability is the fraction of molecules that enter the duct and eventually escape through the outlet at some future time. Thus, it is necessary to consider the gas molecules that escape through the outlet after any given number of interior wall collisions, or particle moves. The conductance probability is therefore determined by the limit of Ψ_n as the number of particle moves n approaches infinity,

$$\Psi = \lim_{n\to\infty} \Psi_n = \lim_{n\to\infty} \mathbf{e}_{N+2}^T A^n \mathbf{v}^{(0)}. \tag{5.17}$$

It is very important to note that equality in (5.17) is only with respect to the Markov chain representation of the stochastic process, and it does not hold for the true conductance probability of the free molecular duct. The actual motion of the gas molecules is continuous throughout the duct, and is not limited to a finite number of interior states. The Markov chain is only an approximation to the real stochastic process, which, in actuality, is an uncountably infinite-state Markov process. Therefore, there exists a truncation error between the conductance probability of Markov chain simulation in (5.17), that depends on the number N of interior wall states.

The critical computing task of the Markov chain simulation is to effectively capture the expected long-term behavior of the stochastic system in order to yield a consistent estimate of the conductance probability. Specifically, this requires an accurate approximation of the limit $\lim_{n\to\infty} A^n \mathbf{v}^{(0)}$ in (5.17) for the conductance probability Ψ. There are three solution techniques implemented in this section for approximation of this limit: (i) a complete eigensystem decomposition of the tran-

sition probability matrix A; (ii) a successive squaring of the transition probability matrix A; and (iii) a marching technique where the initial probability distribution $\mathbf{v}^{(0)}$is simply multiplied by the matrix A repeatedly. It should be noted that only the solution technique (i) actually attempts to solve for the limit in (5.17) directly. The other two methods adopt an iterative approach and solve for the probability Ψ_n in (5.16) for an increasing number n of interior wall collisions. The iterative approaches (ii) and (iii) continue until the following stopping criterion is reached,

$$\frac{|\Psi_a - \Psi_b|}{\Psi_b} < \epsilon \quad \text{with } a < b, \tag{5.18}$$

at which point the probability Ψ_b is taken as an approximation to the conductance probability Ψ in the Markov chain simulation. In this investigation, $\epsilon = 10^{-14}$ is used for the stopping criterion (5.18) of the iterative Markov chain simulations.[15]

The transition probability matrix A in (5.9), or Markov matrix, is diagonalizable [50, 176]. That is,

$$A = X\Lambda X^{-1},$$

where the matrix X is the complete set of eigenvectors of A (each column of X is a distinct eigenvector), and Λ is a diagonal matrix with the corresponding eigenvalues of X. If the complete eigensystem $X\Lambda X^{-1}$ is known for A, it is then possible to find the following limit exactly,

$$\begin{aligned} \lim_{n\to\infty} A^n &= X\Lambda X^{-1} X\Lambda X^{-1} \cdots \\ &= \lim_{n\to\infty} X\Lambda^n X^{-1}, \end{aligned}$$

when the limit $\lim_{n\to\infty} \Lambda^n$ exists. All the eigenvalues λ on the diagonal of Λ, for the transition probability matrix A in (5.9) are real valued and in the interval $(0, 1]$;

[15]Note that the stopping criterion $\epsilon = 10^{-14}$ is the same criterion adopted for the conjugate gradient solver used in the finite-state linear system simulation presented in Section 5.3.

therefore, the limit $\lim_{n\to\infty}\Lambda^n$ exists. In general, the calculation of the eigensystem of an $N \times N$ matrix A requires $\mathcal{O}(N^2)$ operations to obtain the complete set of eigenvalues, and up to $\mathcal{O}(N^3)$ operations to obtain the complete eigensystem decomposition $A = X\Lambda X^{-1}$. Unfortunately, high-performance eigensystem solvers require sophisticated algorithm development that is beyond the scope of this investigation. Thus, for the first solution technique considered in this section, the mathematical software MATLAB [105] is used in this investigation to solve the complete eigensystem for the transition probability matrix A (5.9). Once the eigensystem $A = X\Lambda X^{-1}$ is found, the exact limit for the Markov chain in (5.17) is calculated for the conductance probability Ψ. Note that the overhead present in MATLAB does not make direct timing comparisons possible for the other solution techniques, which use optimized, problem-specific algorithms compiled in C/C++. It is possible, however, to illustrate the general convergence rate of the eigensystem solution technique for the Markov chain simulation.

It is important to recall that the conductance probability Ψ in (5.17) is only exact for the finite state Markov chain approximation, and not the true continuous stochastic process. As such, even the exact limit for Ψ in (5.17) is still only an approximation, albeit a consistent one, to the true conductance probability of the free molecular duct. Since there exists a truncation error caused by the finite-state approximation of the continuous system, the accuracy of the limit $\lim_{n\to\infty}\Psi_n$ only needs to be within this truncation error to maintain a consistent approximation. The complete eigensystem of the A matrix (5.9) does provide a tremendous amount of detail about the transient behavior of the stochastic system as the Markov chain converges to the long-term system equilibrium. However, this amount of detail is simply not necessary for accurately calculating the conductance probability Ψ of a

free molecular duct. Consequently, the eigensystem solution technique appears, and justifiably so, as a significant amount of unnecessary work for calculating the exact limit for Ψ in (5.17). In general, there are more efficient approaches for approximating the limit $\lim_{n\to\infty} A^n \mathbf{n}\mathbf{v}^{(0)}$, as illustrated by the other two solution techniques.

The second solution technique considered in this section approximates the limit $\lim_{n\to\infty} A^n$ by simply squaring the transition probability matrix A (5.9) repeatedly. That is, the matrix A is first multiplied by itself to obtain A^2, the matrix A^2 is then multiplied by itself to obtain $A^4, \ldots$, and so forth. The result of m successive squarings of the matrix A (*i.e.* $A, A^2, A^4, \ldots, A^{2^m}$) clearly converges exponentially to the limit $\lim_{n\to\infty} A^n$. As a consequence, only a small number of successive squarings are expected to be necessary in order to obtain an accurate approximation of the limit for the conductance probability. By modifying the indices in (5.16), let $\Psi_m = \mathbf{e}_{N+2}^T A^{2^m} \mathbf{v}^{(0)}$ represent the probability a gas molecule in the Markov chain simulation reaches the outlet state within the first 2^m particle moves. To simulate the conductance probability with the Markov chain, the successive squaring of the the A matrix continues until the stopping criterion in (5.18) for Ψ_{m-1} and Ψ_m is reached. Using this stopping criterion, less than 10 successive squarings of the A matrix (5.9) are typically needed to approximate the conductance probability for the free molecular duct lengths considered in this investigation. While the total number of successive squarings is very small, a single multiplication of two $N \times N$ matrices still requires $\mathcal{O}(N^3)$ operations. Therefore, the overall operation count of the second solution technique for the Markov chain simulation is $\mathcal{O}(N^3)$ as well.

The third solution technique, referred to as the marching method, approximates the limit $\lim_{n\to\infty} A^n \mathbf{v}^{(0)}$ by directly calculating each update of the probability distribution vector $(\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots)$ using (5.8). After each calculation of probability

distribution $\mathbf{v}^{(n)}$ for $n = 1, 2, \ldots$, the approximation Ψ_n (5.16) to the conductance probability is found. The Markov chain simulation using the marching method continues until successive approximations, Ψ_{n-1} and Ψ_n, of the conductance probability satisfy the stopping criterion in (5.18). The primary cost of the marching method is the matrix-vector calculation performed in (5.8) for the update of the probability distribution $\mathbf{v}^{(n)}$, which requires $\mathcal{O}(N^2)$ operations. The overall cost of the Markov chain simulation using the marching method, however, depends on the total number of successive approximations of Ψ_n (5.16) needed until the stopping criterion is reached. Let s denote the smallest number of particle moves necessary for Ψ_{s-1} and Ψ_s to reach the stopping value $\epsilon = 10^{-14}$ in (5.18). If λ_1 denotes the largest eigenvalue of the A matrix (5.9) less than one, then s can be estimated by the following,

$$s \approx \frac{\log \epsilon}{\log \lambda_1}. \tag{5.19}$$

The eigenvalue λ_1 approaches a constant value as the number of interior wall states N increases; and thus, the total number of particle moves needed to reach the stopping criterion s for the marching method approaches a constant value as well.[16] Therefore, the overall operation count using the marching method remains $\mathcal{O}(N^2)$, which is asymptotically the lowest computational cost of the three solution techniques for the Markov chain simulation.

The conductance probability Ψ is found for a free molecular duct with a length to height ratio $L = 2$ using the three different solution techniques for the Markov chain simulation. More specifically, the Markov chain simulation is performed using systems with $8 \leq N \leq 8192$ interior wall states (except for the MATLAB solution of

[16]It is interesting to note that the largest eigenvalue $\lambda_1 < 1$ has a strong dependence on the duct length to height ratio L. In particular, λ_1 monotonically tends to one as L increases. For the range of free molecular duct geometries considered in this investigation ($0.5 \leq L \leq 10$), the number of updates needed to reach the stopping criterion s increases from $s = 22$ ($L = 0.5$) to $s = 500$ ($L = 10$).

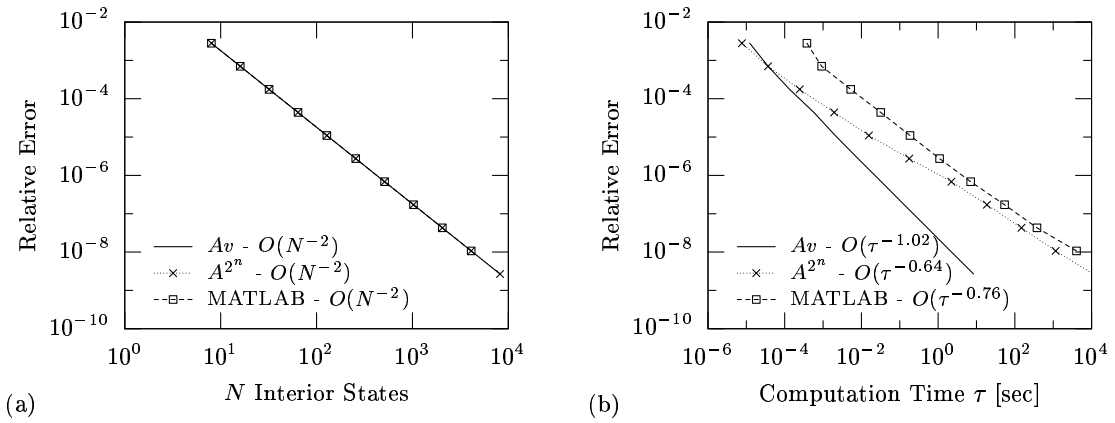

Figure 5.3: Convergence of the relative error in the Markov chain simulation of the conductance probability Ψ (for $L = 2$): (a) error convergence with respect to the number of interior wall states N; and (b) error convergence with respect to the computation time τ.

the eigensystem which only uses up to $N = 4096$ interior states). The convergence of the relative error[17] for these Markov chain simulations is given in Figure 5.3. There is no visual difference between the error convergence of the three different Markov chain simulations with respect to the number of interior wall states N, as illustrated in Figure 5.3(a). This is to be expected because any differences between the approximation of Ψ using three solution techniques should be on the order of the stopping value $\epsilon = 10^{-14}$ in (5.18). Furthermore, the relative error of the Markov chain simulation in Figure 5.3(a) clearly demonstrates quadratic convergence with respect to the number of interior wall states N. That is to say, if the number of interior wall states N is doubled, then the accuracy of the conductance probability Ψ found by the Markov chain simulation is quadrupled.

Unlike the error convergence with respect to N, the error convergence of the three different solution techniques is noticeably different with respect to the computation

[17]The relative error is the difference between the simulation solution and the exact solution normalized by the exact solution. Here the "exact" solution is taken from the more accurate Nyström method which is shown in Section 5.4 to have a stable relative error less than 10^{-12} for the duct lengths under consideration.

time τ, as illustrated in Figure 5.3(b). Not surprisingly, the Markov chain simulation using MATLAB to calculate the complete eigensystem of the A matrix (5.9) is the slowest Markov chain simulation, regardless of the desired accuracy. The error convergence of the MATLAB method with respect to time is $\mathcal{O}\left(\tau^{-0.76}\right)$, which indicates that the algorithm MATLAB uses to solve the eigensystem with N interior states requires approximately $\mathcal{O}\left(N^{2.63}\right)$ operations. The successive squaring technique is the next fastest Markov chain simulation (except for the $N = 8$ case). The error convergence of the successive squaring with respect to time is $\mathcal{O}\left(\tau^{-0.64}\right)$ because the matrix multiplication process is $\mathcal{O}\left(N^3\right)$; and as a consequence, the method has the slowest asymptotic convergence rate of the three methods in terms of computation speed. The marching method is consistently the fastest Markov chain simulation to reach almost all the error levels tested; and as anticipated, its convergence with respect to time is approximately linear. As an example of its speed, the marching method is 8 times faster reaching an error level of 10^{-5} ($N = 128$) than the successive squaring technique; and over 90 times faster than the MATLAB method. It should be noted that in cases when the number of updates needed by marching method to reach the stopping criterion (5.18) is greater than the number of interior wall states N the operation count is really $\mathcal{O}\left(N^3\right)$. Therefore, it is possible for the successive squares technique to actually be faster in some instances as illustrated in Figure 5.3(b).

5.3 Finite State Linear System Simulation

It is possible, using the same finite state representation of the free molecular duct as the Markov simulation in Section 5.2, to construct an even faster method for calculating the conductance probability. To achieve this improved performance one must

slightly alter the physical viewpoint of the Markov chain simulation. Rather than tracking the evolution of $\mathbf{v}^{(n)}$ in (5.8) for the expected distribution of the molecules after n moves, suppose one knew the probability that a molecule at a specific interior wall state would eventually escape the duct through the outlet. Let f_i denote the probability a gas molecule in the interior wall state i escapes the outlet after any number of moves (possibly infinite). If the vector $\mathbf{f} = (f_1, \ldots, f_N)$ is known, then the conductance probability can be calculated from the known quantities $\mathbf{b}$ (5.14) and $\rho(L)$ (5.15) used previously in the Markov chain simulation. That is,

$$\Psi = \mathbf{f} \cdot \mathbf{b} + \rho(L). \tag{5.20}$$

Again, it is important to note that the result in (5.20) is exact with respect to the finite state stochastic process, but not the true continuous process. The solution in (5.20), however, serves as a consistent approximation of the true stochastic process of free molecular duct flow to within a truncation error, which decreases as the number N of interior wall states increases.

The probability f_i of a molecule eventually escaping the duct from state i through the outlet is equal to the probability the molecule directly escapes the outlet on its next move, plus the probability it jumps to any other interior wall state and eventually escapes through the outlet from that state. Accordingly, using the previously derived vector and matrix quantities $\mathbf{b}$ (5.14) and K (5.11) for the Markov chain simulation, an implicit formula can be given by

$$f_i = h_i + \sum_{j=1}^{N} K_{ji} f_j,$$

which in matrix-vector notation becomes

$$\mathbf{f} = \mathbf{h} + K^T \mathbf{f}. \tag{5.21}$$

The vector $\mathbf{f}$, for the eventual outlet escape probability, is then found by rearranging the linear system in (5.21) to yield

$$\mathbf{f} = -(K - I)^{-1}\mathbf{h}, \tag{5.22}$$

where I represents the $N \times N$ identity matrix. Recall from Section 5.2 that the matrix K is symmetric; thus, its transpose can be dropped from the result in (5.22). The solution to the linear system (5.22) exists for all finite duct lengths,[18] and can be solved by any one of the many available methods [176, 189, 105, 147, 20]. Once the eventual outlet escape probability $\mathbf{f}$ is known, the conductance probability of the duct is then determined from (5.20).

The linear system solution in (5.22) for the eventual outlet escape probability $\mathbf{f}$ is found using the iterative conjugate gradient method [176], for this investigation. The relative error used as the stopping criterion for the iterative method is taken to be the same as for the Markov simulation in Section 5.2 (*i.e.* $\epsilon = 10^{-14}$). Similar to the Markov chain simulation, the error of the linear system simulation is taken relative to the Nyström method, which is discussed in greater detail in Section 5.4. The error convergence and computation time are given in Figure 5.4 for the linear system simulation of the conductance probability. Specifically, the relative error[19] and the computation time[20] are found for three different duct length to height ratios ($L = 2, 5, 10$), and for a number of interior states N in the range $8 \leq N \leq 2^{16}$.

Both the linear system simulation and the Markov chain simulation both repre-

[18] As the duct length to height ratio approaches infinity, the associate K matrix in the finite state linear system simulation becomes more ill-conditioned and approaches a singular matrix in this limit. However, for duct length to height ratios $L \leq 100$, the common methods for solving linear systems of equations encounter no noticeable stability problems.

[19] The relative error is the difference between the simulation solution and the exact solution normalized by the exact solution. Here the "exact" solution is taken from the more accurate Nyström method which is shown in Section 5.4 to have a stable relative error less than 10^{-12} for the duct lengths under consideration.

[20] The computation time τ is for a single 3.06 GHz Intel Xeon processor.

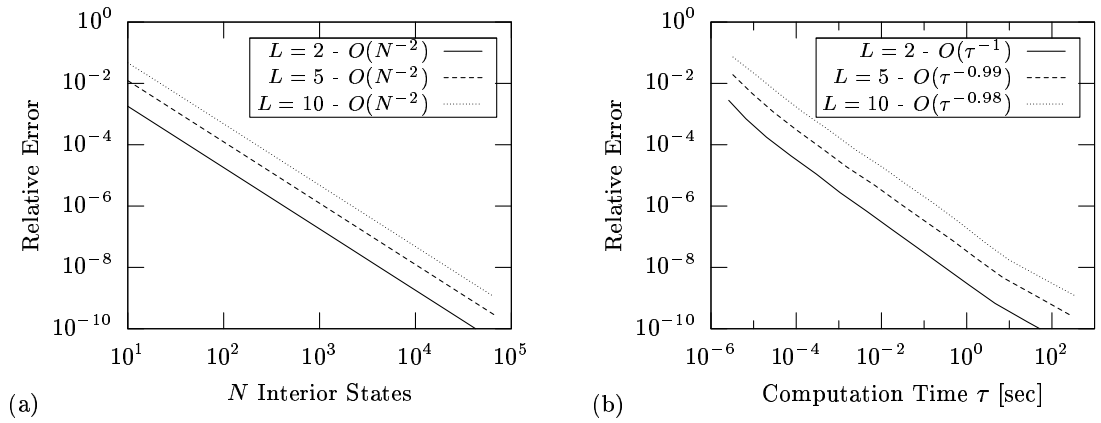

Figure 5.4: Convergence of the relative error in the finite-state linear system simulation of the conductance probability Ψ (for $L = 2$, 5 and 10): (a) convergence with respect to the number of interior duct states N; and (b) convergence with respect to the computation time τ.

sent the same discrete stochastic process. Since both methods use a stopping criterion of $\epsilon = 10^{-14}$, the error convergence (or equivalently, the conductance probability) is expected to appear the same for the $L = 2$ case. As with the Markov chain simulation, the error in Figure 5.4(a) displays quadratic convergence with an increasing number N of interior wall states. Furthermore, given the same number N of interior wall states, the error of the linear system simulation is found to increase with the duct length to height ratio L.

In most cases, including the results from this investigation, the conjugate gradient method for solving linear systems typically requires $\mathcal{O}(N^2)$ operations for an $N \times N$ matrix. Thus, the error convergence relative to the computation time τ is expected to be linear as illustrated in Figure 5.4(b). The $\mathcal{O}(\tau^{-1})$ error convergence rate is the same as achieved by the Markov chain marching simulation of Section 5.2; however, the linear system simulation is approximately 5-7 times faster. Note that the error convergence relative to the computation time slows down as the duct length to height ratio increases. The slower time is not only due to the increase in error seen

in Figure 5.4(a); there is also an additional increase in computation cost attributed to the slower convergence of the conjugate gradient method as the duct length to height ratio increases.

5.4 Nyström method

It is possible to develop an alternate simulation method for the conductance probability by considering what happens to the finite state linear system simulation developed in Section 5.3 when it is extended to the continuous limit. In the continuous limit, the vectors in (5.21) become continuous functions and the matrix-vector operation becomes an integral kernel operator. As a result, the continuous form of (5.21) becomes an implicit integral equation defining an unknown function instead of the eventual outlet escape probability vector $\mathbf{f}$. Specifically, the integral equation of this continuous form is classified as a Fredholm integral equation of the second kind (see [3, 7, 34, 40] for more background on the theory). There exists analytical series solutions to the Fredholm integral equations; however, these series solutions are not always easy to represent in a closed-form. Unfortunately the integral equation resulting from the simulation of the conductance probability falls under this category, meaning no exact closed-form solution is known (at least by the author).

Although an exact solution is elusive for the free molecular conductance probability, there is a very robust numerical technique, referred to as the Nyström method (see [7, 40]), for solving Fredholm integral equations of the second kind. The Nyström method approximates the solution of the integral equation simply by discretizing the integral kernel operator with an appropriately selected numerical integration rule (*e.g.* Newton-Cotes, Gauss-Legendre, *etc.*). The integral equation, once discretized, becomes a well-defined linear system that can be solved for certain points of the

unknown function. The solution procedure is the same as the finite state linear system simulation; thus, the method developed earlier in Section 5.3 is actually an example of the Nyström method. The finite state linear system simulation, however, is a relatively crude implementation of the Nyström method, as there are many numerical integration rules available that offer much greater accuracy than the second-order global accuracy observed in Section 5.3. Specifically in this investigation, the Nyström method is used in conjunction with the Gauss-Legendre integration rules, which are the most accurate for most well-behaved one dimensional integrals.

The resulting Nyström method is, in fact, the most accurate simulation technique developed in this investigation for the free molecular conductance probability. In fact, for certain duct geometries, the Nyström solution appears to be accurate to within machine precision. In addition to its accuracy, the Nyström method is also the fastest simulation technique developed in this investigation, including the QMC particle simulation developed in Section 5.5. The primary computational cost of the Nyström method is the solution of the linear system associated with the discretized integral equation; and, in this investigation, the size of the linear system does not exceed 150 unknowns. As such, solving such small linear systems requires only a trivial amount of time on a modern computer. It is important, however, to remember that the primary goal of developing the QMC particle simulation in this investigation is not to merely obtain the fastest possible simulation. Rather, the goal is to better understand the abilities and limitations of the QMC method when applied to particle simulations. In particular, the focus is to build a foundation from which more general QMC particle simulations can be developed that achieve an error convergence superior to traditional DSMC.

Continuous analogues of the matrix and vector quantities in (5.20) and (5.21)

are required to solve for the free molecular conductance probability Ψ using the Nyström method. As with all the simulation techniques developed in this chapter, the duct geometry is assumed to have a length to height ratio equal to L. Let $K(x, y)$ represent the continuous analogue of the transition probability matrix K given in (5.11). The function $K(x, y)$ then denotes the probability that a particle on the duct wall at x will next collide with the opposite duct wall in the infinitesimal neighborhood of y. The transition probability function $K(x, y)$ is thus determined directly from the parallel transition probability $\mathcal{T}_{\parallel}$ in (5.7) with $\mu = 1$, which is equal to the non-dimensional duct height. That is,

$$\begin{aligned} K(x,y) &= \mathcal{T}_{\parallel}(x,y;1) \\ &= \frac{1}{2\left((x-y)^2+1\right)^{3/2}}. \end{aligned} \tag{5.23}$$

Because the duct geometry is symmetric along the centerline, the transition probability function $K(x, y)$ is the same if the location y is on the opposite duct wall from x, and if y is on the same wall assuming a reflection at the symmetry plane. As with the other simulation techniques in Sections 5.2 and 5.3, the symmetry of the duct geometry allows for simulation to be reduced to a single wall, which is adopted here for the Nyström method.

Similarly, let $h(x)$ represent the continuous analogue of the outlet escape probability vector $\mathbf{h}$ given in (5.13). The function $h(x)$ denotes the probability that a particle on the duct wall at x will escape the duct through the outlet on its next move. The outlet escape probability function $h(x)$ can be calculated in terms of the

perpendicular transition probability $\mathcal{T}_\perp$ in (5.6); that is,

$$\begin{aligned} h(x) &= \int_0^1 \mathcal{T}_\perp(L-x, y)dy \\ &= \frac{1}{2}\left(1 - \frac{L-x}{\sqrt{(L-x)^2+1}}\right). \end{aligned} \tag{5.24}$$

Note that the symmetry condition permits one or both of the duct walls to be simulated using the same outlet escape probability function $h(x)$ in (5.24). Next, let $b(x)$ represent the continuous analogue of the initial probability distribution vector $\mathbf{b}$ given in (5.14). The initial probability distribution function $b(x)$ denotes the probability that a particle entering the duct through the inlet will first collide with the duct wall in the infinitesimal neighborhood of x. The initial probability distribution function $b(x)$ can likewise be calculated in terms of the perpendicular transition probability $\mathcal{T}_\perp$ in (5.6); more specifically,

$$\begin{aligned} b(x) &= \int_0^1 2\mathcal{T}_\perp(x, y)dy \\ &= 1 - \frac{x}{\sqrt{x^2+1}}. \end{aligned} \tag{5.25}$$

Note that the 2 appears in the integral in (5.25) because only one duct wall is simulated in the Nyström method presented here; and the fact that the particle may first collide with either the top or bottom wall must be accounted for. If both duct walls are simulated, then a different form of the initial probability distribution function $b(x)$ (5.25) is needed.

To complete the process of finding continuous analogues of the matrix and vector quantities in (5.20), let $f(x)$ represent the continuous form of the eventual outlet escape probability vector $\mathbf{f}$. The function $f(x)$ denotes the probability a particle located on the duct wall at x will eventually escape the duct through the outlet after any number of wall collisions (possibly infinite). Moreover, the probability

$f(x)$ of a particle eventually escaping the duct from x through the outlet is equal to the probability the particle directly escapes the outlet on its next move, plus the probability it jumps to any other interior wall state and eventually escapes through the outlet from that point. It is therefore possible to define $f(x)$ in terms of the following implicit integral equation,

$$f(x) = h(x) + \int_0^L K(x,y)f(y)dy. \tag{5.26}$$

Note that the implicit integral equation in (5.26) is the continuous analogue of the linear system derived in (5.21) with the matrix-vector operation replaced by the integral kernel operator. Once the eventual escape probability distribution function $f(x)$ is known, the conductance probability Ψ for the free molecular duct is then given by

$$\Psi = \int_0^L b(x)f(x)dx + \rho(L), \tag{5.27}$$

where $\rho(L)$ (5.15) is the probability of a particle directly escaping the duct from the inlet.

The implicit integral equation for $f(x)$ in (5.26) is an example of a linear Fredholm integral equation of the second kind (see [3, 7, 34, 40] for more details regarding the theory behind these integral equations). There exists an analytical solution, referred to as the Neumann series, for the linear Fredholm integral equation of the second kind. To define the Neumann series for the free molecular duct flow, let $f_n(x)$ denote the probability a particle at x will escape through the outlet of the duct after colliding with the wall n or fewer times. It is possible to determine the Neumann series, $f_0(x), f_1(x), \ldots$, explicitly using the following iterative definition,

$$\begin{aligned} f_0(x) &= h(x) \\ f_{n+1}(x) &= h(x) + \int_0^L K(x,y)f_n(y)dy \text{ for } n = 1, 2, \ldots. \end{aligned}$$

Therefore, the solution for $f(x)$ in the integral equation (5.27) is simply the limit of this Neumann series; that is,

$$f(x) = \lim_{n\to\infty} f_n(x). \tag{5.28}$$

The Neumann series (5.28) converges uniformly to $f(x)$ if the function norm $\|K(x,y)\|$ of the integration kernel is strictly less than one. While the Neumann series represents an analytical solution to the linear Fredholm integral equation of the second kind, it should be noted that there is no guarantee the series can be evaluated in terms of known functions. In the case of the free molecular duct flow, the author is unable to find a closed-form for the Neumann series solution of the integral equation in (5.26).

The convergence of the Neumann series ($\|K(x,y)\| < 1$) is a sufficient condition for the Nyström approximation of $f(x)$ in (5.26) to converge to the true function. The particular choice of the function norm does not matter when establishing the convergence of the Neumann series and any of the easy to calculate norms, such as the Frobenius, L_1, or L_∞ norms, may be used. A brief review of function norms can be found in [40], and their treatment is essentially the same as their counterparts for the finite-dimensional matrix and vector norms [176]. For a general integration kernel $\Xi(x,y)$ defined on the function space $\mathscr{L}^2(a,b)$[21], the Frobenius norm $\|\Xi(x,y)\|_F$ is defined by

$$\|\Xi(x,y)\|_F = \left(\int_a^b \int_a^b |\Xi(x,y)|^2 dxdy\right)^{1/2}.$$

The Frobenius norm over the interval $(0,L)$ of the transition probability function $K(x,y)$ (5.23) is then given by

$$\|K(x,y)\|_F = \frac{1}{4}\left(\frac{L^2}{L^2+1} + 3L\tan^{-1}L\right)^{1/2}. \tag{5.29}$$

[21]The function space $\mathscr{L}^2(a,b)$ is the set of all functions $f(x)$ defined on the interval (a,b) that satisfy the condition $\int_a^b |f(x)|^2 dx < \infty$ in the Lebesgue sense (see Chapter 1 of [40] for more details).

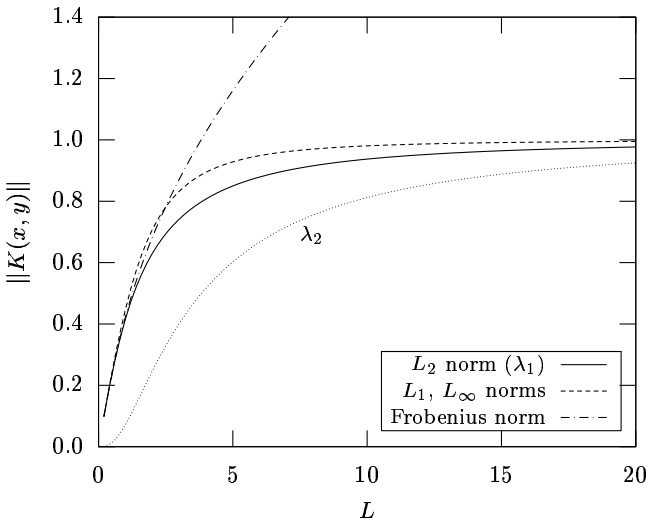

Figure 5.5: The Frobenius, L_1 , L_2, and L_∞ norms of the transition probability function $K(x, y)$ used to solve the conductance probability in a free molecular duct. Note that $\|K(x, y)\|_2 = \lambda_1$, which is the largest eigenvalue of $K(x, y)$, and that the second largest eigenvalue λ_2 is included for reference.

Unfortunately, the Frobenius norm $\|K(x, y)\|_F$ (5.29) is only less than one when $L \leq 3.819$ as shown in Figure 5.5. Based on the Frobenius norm alone, the Nyström method is not guaranteed to yield a consistent approximation of $f(x)$ in (5.26) for all the duct geometries tested in this investigation.

As it turns out, the L_1 and L_∞ norms are better choices for establishing the bound $\|K(x, y)\| < 1$ on the transition probability function (5.23). For a general integration kernel $\Xi(x, y)$ defined on the function space $\mathscr{L}^2(a, b)$, the L_1 norm $\|\Xi(x, y)\|_1$ is defined by

$$\|\Xi(x,y)\|_1 = \sup_{y\in(a,b)} \int_a^b |\Xi(x,y)|dx,$$

and similarly the L_∞ norm $\|\Xi(x, y)\|_\infty$ is defined by

$$\|\Xi(x,y)\|_\infty = \sup_{x\in(a,b)} \int_a^b |\Xi(x,y)|dy.$$

Clearly, the L_1 and L_∞ function norms are the continuous analogues of their respective matrix norms; that is, the L_1 matrix norm is equal to the maximum column

sum of the matrix and the L_∞ matrix norm is equal to the maximum row sum of the matrix [176]. Since the transition probability function $K(x,y)$ (5.23) is symmetric, the L_1 and L_∞ norms of the integration kernel are the same. Specifically,

$$\|K(x,y)\|_1 = \|K(x,y)\|_\infty = \frac{L}{\sqrt{L^2+4}}. \tag{5.30}$$

The result in (5.30) implies that $\|K(x,y)\|_1 = \|K(x,y)\|_\infty < 1$ for all duct geometries $L > 0$, as illustrated in Figure 5.5, thus, the Neumann series (5.28) converges to the eventual outlet escape probability $f(x)$ in (5.26). Most importantly, the Nyström method developed here is guaranteed to yield a consistent approximation to $f(x)$ in (5.27) for the simulation of the free molecular conductance probability Ψ. It is interesting to note that in the limit as $L \to \infty$, the L_1 and L_∞ norms monotonically approach one, which indicates the Neumann series converges more slowly as the duct to length ratio L increases. This seems to suggest that it may become more difficult to obtain an accurate Nyström solution in the limit as well.

While not representable in closed-form, the L_2 function norm offers useful insight into the physical process being simulated, which can, in turn, be exploited by other numerical methods. For any symmetric integration kernel $\Xi(x,y)$ defined on the function space $\mathscr{L}^2(a,b)$, the L_2 function norm is defined as $\|\Xi(x,y)\| = \lambda_1$, where λ_1 is the largest eigenvalue[22] of the kernel. Unfortunately, it is difficult to find a closed form for the eigenvalues of an integration kernel without some *a priori* knowledge of the likely functional form of the associate eigenfunctions. The integral kernel $K(x,y)$ for the transition probability distribution in (5.26) is not exempt from this complication and the eigenvalues of $K(x,y)$ must be determined numerically. Note that if the complete set of eigenvalues and eigenvectors of $K(x,y)$ were known in

[22]If there exists a function $\omega(x) \in \mathscr{L}^2(a,b)$ such that $\int_a^b \Xi(x,y)\omega(y)dy = \lambda\omega(x)$ for all $x \in (a,b)$, then the function $\omega(x)$ is defined as the eigenfunction of the kernel $\Xi(x,y)$ with an associate eigenvalue λ.

closed form, it would be possible to obtain a closed form solution for the Neumann series solution of $f(x)$ in (5.28), and ultimately, the conductance probability Ψ. In this investigation, the L_2 norm $\|K(x,y)\|_2$ is determined by using the classic power iteration [176] to find the largest eigenvalue of the integration kernel $K(x,y)$.[23] This calculation of the L_2 norm $\|K(x,y)\|_2$ using the power iteration is given in Figure 5.5. As with the other function norms, the L_2 norm monotonically increases with the duct to length ratio L (see Figure 5.5); additionally, it appears to be bounded from above the L_1 and L_∞ norms.

The physical significance of the L_2 norm $\|K(x,y)\|_2$ becomes apparent when one considers the probability a particle remains within the duct after a given number of wall collisions. In particular, let $\varphi_n(x)$ denote the probability a particle collides with the duct wall in the infinitesimal neighborhood of x after undergoing n previous wall collisions. It is possible to determine $\varphi_n(x)$ explicitly from the following iterative definition

$$\begin{aligned} \varphi_0(x) &= b(x) \\ \varphi_{n+1}(x) &= \int_0^L K(x,y)\varphi_n(y)dy \text{ for } n = 1, 2, \ldots. \end{aligned} \tag{5.31}$$

Next let $\overline{\varphi}_n$ denote the probability a particle remains within the duct after $(n+1)$ wall collisions; hence,

$$\overline{\varphi_n} = \int_0^L \varphi_n(x)dx. \tag{5.32}$$

The probability $\overline{\varphi}_n$ of a particle remaining within the duct after $(n+1)$ wall collisions is then calculated[24] in Figure 5.6(a) for three different duct geometries ($L = 2, 5, 10$).

[23] The power iteration in [176] is only developed for finding the largest eigenvalue of a symmetric, finite-dimensional matrix. It is possible to adapt the matrix power iteration algorithm into a form for symmetric integration kernels simply by replacing the matrix-vector operations with their appropriate continuous integral analogues. All the integrals in the power iteration performed here are approximated using an 80-point Gauss-Legendre integration rule.

[24] The integrals in (5.31) and (5.32) necessary for calculating $\overline{\varphi}_n$ are approximated numerically using an 80-point Gauss-Legendre integration rule, as with the power iteration.

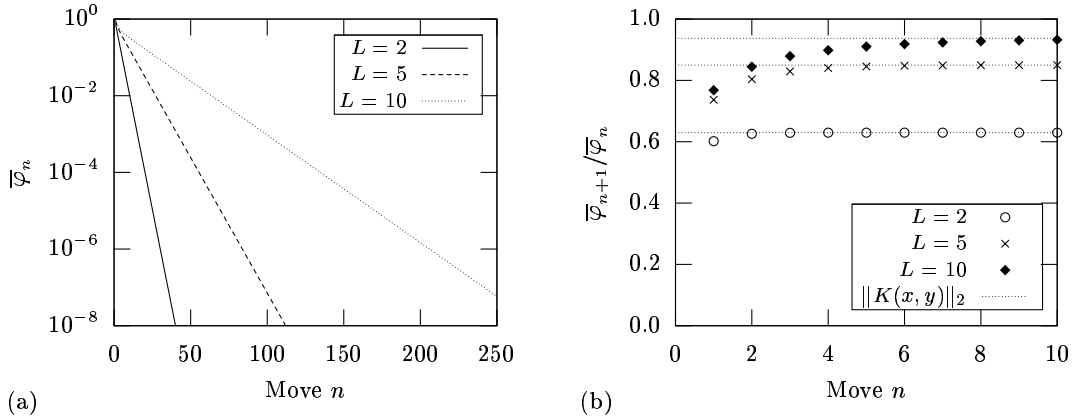

Figure 5.6: The probability $\overline{\varphi}_n$ of a particle remaining within the duct after $n+1$ wall collisions (for $L = 2$, 5 and 10): (a) convergence of $\overline{\varphi}_n$ to zero as the number of wall collisions increases; and (b) convergence of the successive ratio $\overline{\varphi}_{n+1}/\overline{\varphi_n}$ to a constant value.

The convergence of $\overline{\varphi}_n$, for all duct geometries tested, appears to be linear on the semi-log scale plot. As a result, the probability $\overline{\varphi}_n$ of a particle remaining within the duct after $(n+1)$ wall collisions is closely approximated by an equation of the form

$$\overline{\varphi}_n \propto k^n, \tag{5.33}$$

where k is some constant that depends on the length to height ratio L.

In order to approximate k, the successive ratio $\overline{\varphi}_{n+1}/\overline{\varphi_n}$ is found in Figure 5.6(b). That the successive ratio appears to converge to a constant value k suggests the following process physically occurs in free molecular duct flow regardless of the simulation technique used to solve it. Given a population of particles entering the free molecular duct, roughly the same fraction k of the remaining population is likely to stay within it after each subsequent wall collision.[25] Conversely, the fraction $(1-k)$ is likely to escape the duct from the remaining population after each subsequent wall collision. Based on inspection, the value of this constant k is equal to the L_2 norm $\|K(x,y)\|_2$ as illustrated in Figure 5.6(b). Therefore, the average fraction of particle

[25]This is excluding the first several wall collisions where the initial probability distribution exerts some effect on the probability of escaping the duct back through the inlet.

population beginning inside the duct that still remains inside after n moves is closely approximated by $\|K(x,y)\|_2^n$. This result serves as a useful estimate in determining the appropriate number of moves in the sample trajectories generated for the absorption weighted Monte Carlo and quasi-Monte Carlo simulations (developed in Section 5.5).

It is not surprising that the constant k in (5.33) is equal to $\|K(x,y)\|_2$, because the power iteration algorithm for calculating the largest eigenvalue of an integration kernel is very similar to the calculation of $\overline{\varphi}_{n+1}/\overline{\varphi_n}$ using (5.31) and (5.32). The ratio λ_2/λ_1 of the largest and second largest eigenvalues affects the rate at which the power iteration converges to the eigenvalue; specifically, the convergence slows as $\lambda_2/\lambda_1 \to 1$. Based on the value of λ_2 (see Figure 5.5) for the transition probability $K(x,y)$, the same behavior is observed in Figure 5.6(b) for the successive ratio $\overline{\varphi}_{n+1}/\overline{\varphi_n}$.

The Nyström method solves the linear Fredholm integral equation of the second kind by suitably discretizing the integral kernel operator and then solving the resulting linear system directly, as done previously in Section 5.3. Let $\mathscr{I}_n(a,b) = \{\mathbf{x}, \mathbf{w}\}$ denote an n-point numerical integration rule with sample points $\mathbf{x} = (x_1, \ldots, x_n) \in (a,b)$ and sample weights $\mathbf{w} = (w_1, \ldots, w_n)$ subject to the constraint $\sum_{i=1}^n w_i = (b-a)^{-1}$. Given any function $\phi(u)$ defined on the interval $u \in [a,b]$, the n-point integration rule then approximates the integral of $\phi(u)$ by the following weighted average of the function samples,

$$\int_a^b \phi(u)du \approx \sum_{i=1}^n w_i \phi(x_i).$$

The Nyström method using $\mathscr{I}_n(0,L)$ reduces the integral equation for the eventual outlet escape probability $f(x)$ in (5.26) to the following system of n linear equations,

$$f(x_i) = h(x_i) + \sum_{j=1}^n w_j K(x_i, x_j) f(x_j) \text{ for } i = 1, \ldots, n, \tag{5.34}$$

in n unknowns, $f(x_1), \ldots, f(x_n)$. Any linear system solver can then be used to calculate the unknowns $f(x_1), \ldots, f(x_n)$ in (5.34); in this investigation, the standard Gaussian elimination with partial pivoting [176] is adopted. While there exists linear system solvers with better asymptotic performance than Gaussian elimination, here the maximum number of points n in the integration rule is 150. Therefore, solving the linear system in (5.34) with Gaussian elimination requires a minimal amount of time on any modern computer.

The Nyström method does not solve the eventual outlet escape probability $f(x)$ in (5.26) completely. Instead, the Nyström method approximates specific values of the function $f(x_i)$ at locations corresponding to the sample points $\mathbf{x}$ in the integration rule $\mathscr{I}_n(0, L)$. That the Nyström method only solves for $f(x)$ at a limited number of locations is of no consequence when calculating the conductance probability Ψ of the free molecular duct flow. To calculate the conductance probability Ψ, the integral of $f(x)$ must also be approximated in (5.27), which corresponds perfectly with respect to the Nyström method since $f(x)$ is only known at the sample points of the integration rule. Therefore, the Nyström method using the integration rule $\mathscr{I}_n(0, L)$ yields the following estimate of the conductance probability (5.27),

$$\Psi = \rho(L) + \sum_{i=1}^{n} w_i f(x_i) b(x_i), \tag{5.35}$$

where $\rho(L)$ (5.15) is the probability of a particle directly escaping the duct from the inlet and $b(x)$ (5.25) is the initial probability distribution function. In this investigation, the numerical integration rule $\mathscr{I}_n(0, L)$ of the Nyström method is chosen to be the n-point Gauss-Legendre integration rule appropriately scaled to the interval $[0, L]$ (see [1, 20] for tables of $\{\mathbf{x}, \mathbf{w}\}$). The Gauss-Legendre rules are perhaps the most accurate numerical approximation to the integral of a well-behaved, one di-

mensional function over a finite interval, which is why they are selected here for the Nyström approximation (5.35) to the conductance probability.

Despite the best efforts to find the set of eigenfunctions for the transition probability function $K(x, y)$ (5.26), they remain elusive to the author; thus, no analytical series solution to the conductance probability Ψ is available. Given that the Nyström method is the most accurate simulation technique presented in this chapter, its accuracy can only be verified by comparing the solution to itself. The Nyström method described in (5.34) and (5.35) is solved using every n-point Gauss-Legendre integration rule in the range $4 \leq n \leq 150$, to determine the best possible approximation to the exact value of the conductance probability Ψ. The Nyström solutions for 10 consecutive integration rules are then averaged together to determine the sample mean and sample variance.[26] The sample mean which is found to have the lowest sample variance is then taken to be the best possible approximation of the conductance probability by the Nyström method. The motivation for adopting this strategy is based on the convergence pattern of the Nyström method observed when the number n of Gauss-Legendre points increases in Figure 5.7. In particular, the relative error found in Figure 5.7 is the normalized difference between the Nyström solution using a specific n-point Gauss-Legendre rule and the best possible approximation determined from the sample statistics. The error convergence of the Nyström solution to the conductance probability follows the same basic pattern for all the free molecular duct geometries tested in this investigation. There are three parts to the convergence pattern illustrated in Figure 5.7: (i) rapid initial convergence of the Nyström solution to a minimum error level; (ii) stabilization around this minimum error level for at least the next 10 Gauss-Legendre rules; and (iii) eventual divergence of the Nyström

[26]That is, the solution mean and variance are taken by averaging the result of the Nyström method using the $n, n+1, \ldots, n+9$-point integration rules, for $4 \leq n \leq 141$.

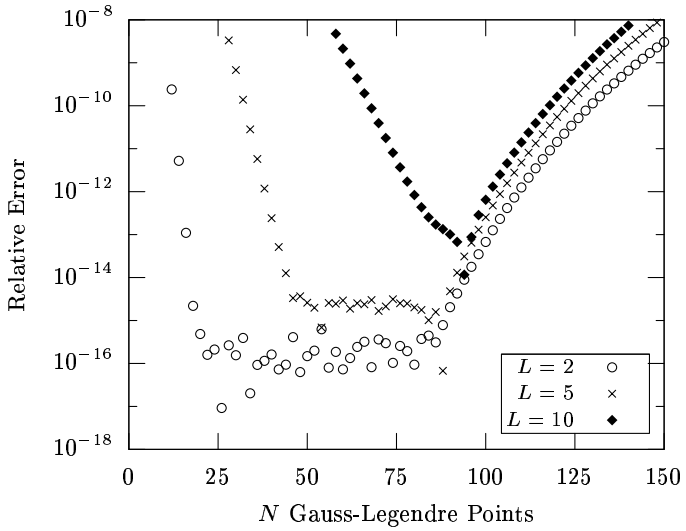

Figure 5.7: Error convergence of the Nyström method using an n-point Gauss-Legendre rule to solve the conductance probability Ψ.

solution as a result of the accumulation of round-off error.

The Nyström solution for the conductance probability of the $L = 2$ duct geometry achieves a minimum relative error of approximately 10^{-16} for all the n-point Gauss Legendre rules in the range $21 \leq n \leq 80$, as shown in Figure 5.7.[27] The accuracy and speed of the Nyström method clearly makes it the best simulation for the conductance probability. For example, using the finite state linear system with the much cruder integral approximation (see Section 5.3), requires $N = 2^{16}$ interior states to be simulated in order to reach the same accuracy as the Nyström method using a 13-point integration rule. Since both methods obtain their approximation in part by solving a linear system, the difference in computation time between the two methods is tremendous (*i.e.* solving a $2^{16} \times 2^{16}$ system versus a 13×13 system). It appears in Figure 5.7 that the accuracy of the Nyström method decreases as the duct length to height ratio L increases. In addition, the number of Gauss-Legendre rules that produce a solution near the minimum error observed in the stabilization region

[27]Note that this error level is at or near the machine precision.

decreases. More specifically, the error in the Nyström solution for the $L = 5$ duct geometry is slightly higher, achieving a minimum error of approximately $2 \cdot 10^{-15}$ for all the n-point Gauss Legendre rules in the range $49 \leq n \leq 88$. The error in the Nyström solution for the $L = 10$ duct geometry is higher still, achieving a minimum error of approximately 10^{-13} for all the n-point Gauss Legendre rules in the range $84 \leq n \leq 98$. Of all the integration rules, the 80-point Gauss-Legendre rule is found to consistently yield one of the best approximations of the conductance probability using the Nyström method. In fact, the relative error of the Nyström method using the 80-point Gauss-Legendre rule is less than 10^{-12} for all the duct geometries in the range tested in this investigation ($0.5 \leq L \leq 10$).

While there is no exact solution available to compare with the Nyström method, it is possible to check the numerical solution obtained in this investigation against the approximate solution of Clausing [30]. The solution in [30] is obtained by assuming the eventual outlet escape probability $f(x)$ in (5.26) is a linear function in x. Given the symmetry of the duct, the assumed linear form of $f(x)$ can be represented in terms of a single free parameter $\alpha(L)$, which is permitted to vary with the duct length to height ratio L. In [30], the free parameter $\alpha(L)$ is calculated differently for the wide duct regime $L < 1$ and the narrow duct regime $L > 1$. The form of $\alpha(L)$ in the wide duct regime ($L < 1$) is designed to yield the correct solution in the limit $L \to 0$. Similarly, the form of $\alpha(L)$ in the narrow duct regime ($L > 1$) is intended to produce the correct asymptotic convergence of the conductance probability (*i.e.* $\Psi = \mathcal{O}\left(L^{-1} \log L\right)$) in the limit $L \to \infty$. The approximate solution of the conductance probability given by Clausing [30] is in agreement with the Nyström method developed in this investigation, as illustrated in Figure 5.8. Specifically, the difference between the two solutions is negligible (*i.e.* less than 0.06%) in the wide

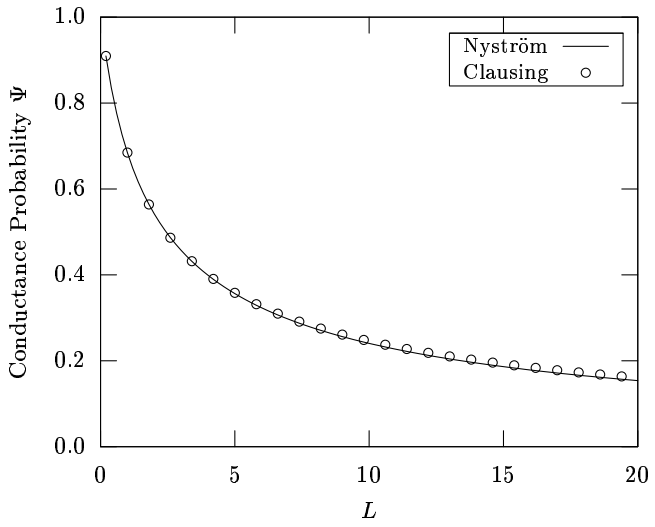

Figure 5.8: Comparison of the conductance probability Ψ calculated from the Nyström method and the approximation of Clausing.

duct regime of $L < 1$. The difference, however, appears to grow steadily as the duct becomes more narrow reaching a difference of more than 4% for the $L = 20$ duct geometry.

5.5 Particle Methods

There are three particle simulations tested in this investigation for free molecular flow in a two dimensional duct: (i) the traditional test particle Monte Carlo simulation; (ii) the absorption weighted Monte Carlo (AWMC) particle simulation; and (iii) the quasi-Monte Carlo (QMC) particle simulation. In Section 5.5.1, the implementation of the test particle Monte Carlo simulation is discussed, and the convergence of the method is demonstrated for a duct geometry with a length to height ratio $L = 2$. Two attempts are also made to convert the test particle simulation directly into a computationally efficient QMC simulation by simply replacing the pseudo-random number generator with a low-discrepancy sequence. These initial attempts, however, do not produce a QMC particle simulation with a near-linear error convergence as

intended. In Section 5.5.2, the implementation of the AWMC simulation is reviewed and convergence is shown for the same duct geometry as the test particle Monte Carlo simulation. Absorption weighting is a variance reduction technique that is commonly used in the Monte Carlo simulation of radiation transport. Because the variance of the AWMC simulation is lower than the test particle Monte Carlo simulation, the AWMC simulation has a lower relative error for the same number of samples. The absorption weighting technique also provides an alternate form for the particle simulation that is more amenable to QMC integration than the traditional test particle formulation, based on the results of Sarkar and Prasad [153]. The QMC particle simulation is therefore developed using the absorption weighting technique in Section 5.5.3, and near-linear convergence is demonstrated for the same test case as the test particle Monte Carlo simulation. Most importantly, this section establishes that it is possible to construct a QMC particle simulation of free molecular flow that achieves the theoretical near-linear error convergence rate.

5.5.1 Test Particle Monte Carlo Method

The test particle Monte Carlo simulation is one of the earliest applications of the Monte Carlo method to fluid flows. Davis [36], in particular, develops the test particle Monte Carlo method for free molecular flow through several different pipe and duct geometries. Based on the definition of Hammersley and Handscomb [58], the test particle method is an example of direct simulation Monte Carlo; that is, the stochastic events being simulated correspond directly to the real-life physical processes they represent. Specifically, the method simulates the actual trajectory path of the gas molecules (also referred to here as simply particles) as they travel through the duct under the conditions of free molecular flow. As a consequence,

the test particle Monte Carlo simulation is, perhaps, the most intuitive approach to approximating the conductance probability of a free molecular duct.

As noted, the conductance probability Ψ of a duct is defined as the fraction of particles that enter the duct at the inlet and then eventually escape the duct through the outlet. To obtain an approximation of Ψ using direct simulation, many randomly generated particle trajectories are collected to produce a physically accurate representation of the free molecular flow through the duct. There are two basic random events that occur during a particle trajectory: (i) the initial entry of the particle into the duct through the inlet; and (ii) the particle collisions with the fully diffuse walls of the duct. The probabilistic outcomes of both of these events can be described in terms of the cosine distribution of trajectory angles given in (5.5). It is more convenient, however, to work with these probabilistic outcomes in terms of the location along the wall of the next particle collision instead of the trajectory angle. The probability distribution functions that govern the next location where a particle collides with the wall can be calculated using the perpendicular $\mathcal{T}_\perp$ (5.6) and parallel $\mathcal{T}_\parallel$ (5.7) transition probabilities defined in Section 5.1. The test particle Monte Carlo simulation then tracks these physically accurate particle trajectories from the point of entry into the duct until the particle escapes the interior. While a single random particle trajectory offers little about the expected behavior of the free molecular flow, the collected average of a large number of random trajectories is able to provide a tremendous amount of information. The conductance probability Ψ, which is of specific interest to this investigation, is approximated by simply counting the fraction of the sample trajectories that result in a particle escaping the duct through the outlet. If more sample trajectories are used in test particle Monte Carlo simulation, then naturally the approximation of Ψ is expected to be more accurate.

To proceed more formally, for a duct with a length to height ratio L define

$$T^{(n)} = \left\{ y_0^{(n)}, z_1^{(n)}, z_2^{(n)}, \ldots, z_e^{(n)} \right\} \tag{5.36}$$

as the n^{th} particle trajectory, where $y_0^{(n)} \in [0,1]$ is the point on the inlet plane where the particle enters the duct, and $z_1^{(n)}, \ldots, z_{e-1}^{(n)} \in [0, L]$ are points along the interior walls where the particles collide. The last point $z_e^{(n)}$ in the trajectory $T^{(n)}$ indicates when the particle escapes the duct, and as such, $z_e^{(n)}$ is the only wall location not in the interval $[0, L]$. Since it is possible to have a particle trajectory with an infinite number of wall collisions, the index e of the last point in the trajectory can be any positive integer. The point $z_e^{(n)}$ represents the location where an imaginary wall collision would occur if the duct is assumed to be of infinite length. Two notes need to be stated about the particle trajectory $T^{(n)}$ as it is used for this investigation: (i) if it is clear from the context of the discussion that only a single trajectory is being considered, then the superscript (n) is omitted for convenience; and (ii) the positions of the particle trajectory are represented in terms of a non-dimensional length normalized by the duct height. Each trajectory is given a score $S_{mc}(T^{(n)})$ to determine its contribution to the estimate of the conductance probability. Since the conductance probability is approximated by counting the fraction of sample trajectories that result in the particle escaping the duct through the outlet, the trajectory score $S_{mc}(T^{(n)})$ is simply an indicator function. Specifically,

$$S_{mc}(T^{(n)}) = \begin{cases} 1 & \text{if } z_e^{(n)} > L \\ 0 & \text{if } z_e^{(n)} < 0. \end{cases} \tag{5.37}$$

The test particle simulation then yields the following approximation to the conductance probability Ψ given by

$$\Psi = \frac{1}{N} \sum_{n=1}^{N} S_{mc}(T^{(n)}), \tag{5.38}$$

where N is the total number of sample trajectories simulated.

It is important to understand that the particle trajectories are independent of each other in the estimate of conductance probability (5.38). Therefore, the trajectories may be generated in any order, simultaneously or serially, without affecting the accuracy of the test particle simulation. This is not surprising because free molecular flow is a mathematical approximation to a flow regime where particle-particle collisions are exceedingly rare; thus, the particle trajectories should appear independent in the simulation. Serial generation of the sample particle trajectories is typically the fastest computation approach (see Bird [16]), and is the approach adopted in this investigation. There is no difference in the total number of mathematical operations between the serial and simultaneous generation of the sample trajectories. There is, however, a major difference in the overall memory required during the execution of the algorithm. As previously noted, a particle that undergoes a collision with a diffuse wall loses all memory of it previous trajectory. Thus, at any given point during the generation of the sample trajectory, only two positions ever need to be stored in memory: (i) the current position of the particle on the boundary of the duct; and (ii) the future position of the next wall collision. If the sample trajectories are generated simultaneously rather than serially, significantly more memory is necessary to store each pair of locations for every trajectory being simulated. As such, the memory latency ultimately determines the computational cost difference between the serial and simultaneous generation of the sample trajectories. The memory required for serial generation is typically so small that the complete algorithm is able to be executed from the processor chip cache of most modern desktop computers. In contrast, the simultaneous generation of the sample trajectories typically requires a large number of calls to the computer's RAM; especially, as in this investigation, when more than

a million samples are needed. Consequently, the simultaneous generation of the sample trajectories runs significantly slower as the memory latency is roughly an order of magnitude greater for accessing the RAM than the chip cache.

The first point y_0 of a trajectory T corresponds to the point where the particle enters the interior of the duct from the inlet. As stated in the initial problem description (see Section 5.1), the inlet of the duct is attached to a reservoir of infinite expanse, and the gas molecules in the reservoir are assumed to be in local thermodynamic equilibrium with zero drift velocity. These assumptions imply that a particle is equally likely to cross the inlet plane at any point $y_0 \in [0, 1]$ Therefore, $y_0 = u$, for each trajectory generated in the test particle simulation, where $u \in \mathscr{U}(0, 1)$ is a uniformly distributed, random variate in the interval $[0, 1]$.

The second point z_1 of a trajectory T corresponds to the point where the particle first collides into the interior duct wall. Or, if the particle escapes the duct directly, z_1 is the imaginary location the particle would collide with the wall if the duct was of infinite length. To include both possibilities, the interior duct wall and its imaginary extension are collectively referred to as the wall plane. Let $b'(y_0, z_1)$ denote the probability a particle on the inlet plate at y_0 first intersects either the top or bottom wall plane at z_1. It is possible to define $b'(y_0, z_1)$ in terms of the perpendicular transition probability $\mathcal{T}_\perp$ (5.6) as follows,

$$\begin{aligned} b'(y_0, z_1) &= \mathcal{T}_\perp(y_0, z_1) + \mathcal{T}_\perp(1 - y_0, z_1) \\ &= \frac{y_0 z_1}{2\left(y_0^2 + z_1^2\right)^{3/2}} + \frac{(1 - y_0) z_1}{2\left((1 - y_0)^2 + z_1^2\right)^{3/2}}. \end{aligned} \tag{5.39}$$

Note that z_1 is simply the downstream distance from the inlet regardless of the wall surface. Both the probability of intersecting with the lower wall plane (first term) and the probability of intersecting with the upper wall plane (second term)

are included in the distribution (5.39). This first intersection with the wall plane is the only time during the test particle simulation that the distinction between the upper and lower walls is made. There is no difference in the treatment of the walls for all the subsequent particle moves, $z_2, \ldots, z_e$, because of the symmetry of the duct and the symmetry of the parallel transition probability $\mathcal{T}_{\parallel}$ in (5.7). To generate a sample point z_1 for a trajectory, the inverse cumulative transform method (see [47]) is applied to the distribution function $b'(y_0, z_1)$ in (5.39). It is difficult to obtain an explicit inverse of the integral of $b'(y_0, z_1)$; however, it is possible to obtain the inverses of the integrals of each term in (5.39) separately. Each term in (5.39) corresponds to a particle moving toward either the upper or lower wall plane, and both events have an equal 50% probability of occurring. It is, therefore, natural to divide the generation of the sample point z_1 into these two type of events. Let $\mathcal{B}'^{-1}(u; y_0)$ denote the inverse cumulative distribution function of $b'(y_0, z_1)$ in (5.39). Hence, the first trajectory location z_1 along the duct wall is generated by

$$z_1 = \mathcal{B}'^{-1}(u; y_0) = \begin{cases} 2y_0 \frac{\sqrt{u(1-u)}}{1-2u} & \text{if } u < \frac{1}{2} \\ 2(y_0 - 1)\frac{\sqrt{u(1-u)}}{1-2u} & \text{if } u > \frac{1}{2}, \end{cases} \tag{5.40}$$

where $u \in \mathscr{U}(0,1)$ is a uniformly distributed, random variate in the interval $[0,1]\backslash\{\frac{1}{2}\}$. Note that the case $u = \frac{1}{2}$ produces a particle moving exactly parallel to the duct wall (*i.e.* $z_1 \to \infty$), and must be handled accordingly if the specific source used for the random variates contains this value. After z_1 is generated for each trajectory in the test particle simulation, one must check if the particle has escaped the duct (*i.e.* $z_1 > L$). If the particle has escaped, then the trajectory is terminated ($e = 1$), its score $S_{mc}(T)$ is added to the running average approximation in (5.38) for the conductance probability, and a new trajectory is started (as necessary for greater accuracy).

If the particle does not directly escape from the inlet, additional trajectory locations $z_2, \ldots, z_e$ are generated until it eventually escapes at z_e with $e \geq 2$. These locations in the trajectory only involve particle moves between the two parallel wall planes; as such, they are governed by the same probability distribution used for the integral kernel $K(x, y)$ in the Nyström method of Section 5.4. It is simpler, with respect to the inverse cumulative distribution, to represent the transition probability $K(z_i, z_{i+1})$ (5.23) in terms of the difference $\Delta z = z_{i+1} - z_i$ between successive trajectory locations. The probability $K(\Delta z)$ a particle leaving the wall after a diffuse collision will strike the opposite wall a distance Δz from the original location is then given by

$$K(\Delta z) = \frac{1}{2\big((\Delta z)^2 + 1\big)^{3/2}}. \tag{5.41}$$

Each new location of the trajectory, $z_2, \ldots, z_e$, is then generated from the inverse cumulative transform of (5.41) defined as $\mathcal{K}^{-1}(u)$. Hence,

$$\begin{aligned} z_{i+1} &= z_i + \mathcal{K}^{-1}(u) \\ &= z_i + \frac{u - \frac{1}{2}}{\sqrt{u(1-u)}}, \text{ for } 2 \leq i \leq e, \end{aligned} \tag{5.42}$$

where $u \in \mathscr{U}(0, 1)$ is a uniformly distributed, random variate in the interval $(0, 1)$. Note that the cases $u = 0$ and $u = 1$ correspond to $\Delta z \to -\infty$ (escape through the inlet), and $\Delta z \to \infty$ (escape through the outlet), respectively. If the specific source for the random variates includes these values, they must be handled accordingly.

After each new trajectory location z_i (with $2 \leq i \leq e$) is generated in (5.42), one must check if the particle has escaped the interior of the duct. If $z_i < 0$ or $z_i > L$, then, by definition, $i = e$ and the sample trajectory is terminated. The trajectory score $S_{mc}(T)$ is then added to the running average approximation in (5.38) for the conductance probability and a new trajectory is started (as necessary). The test

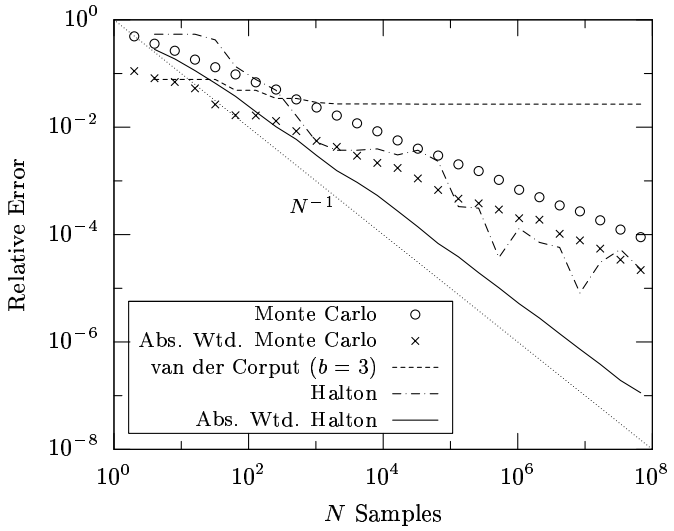

Figure 5.9: The relative error of different particle simulations of the conductance probability Ψ (for $L = 2$).

particle simulation is now completely defined by these three basic steps for producing a sample trajectory in (5.36): (i) generating the inlet plane location y_0 from $\mathscr{U}(0,1)$; (ii) generating the first intersection with the wall plane z_1 using (5.40); and (iii) generating any necessary subsequent moves $z_2, \ldots, z_e$ between the interior walls of the duct using (5.42).

In the preceding discussion of the test particle simulation, no method is prescribed for generating any of the random variates $u \in \mathscr{U}(0,1)$ needed. The test particle Monte Carlo simulation is obtained when the random variates $u \in \mathscr{U}(0,1)$ are produced from a pseudo-random number (PRN) generator appropriately scaled to the unit interval. Figure 5.9 demonstrates that the relative error[28] in the conductance probability Ψ does indeed converge for the test particle Monte Carlo simulation of a free molecular duct with a length to height ratio $L = 2$. The convergence rate of the test particle Monte Carlo simulation is $\mathcal{O}\left(N^{-1/2}\right)$ (where N is the number of sam-

[28]The relative error is the difference between the simulation solution and the exact solution normalized by the exact solution. Here the "exact" solution is taken from the more accurate Nyström method which is shown in Section 5.4 to have a stable relative error less than 10^{-12} for the duct lengths under consideration.

ple trajectories generated), which is the expected rate of all Monte Carlo methods. In general, Monte Carlo methods usually have a noticeable amount of fluctuations present in their error convergence results, and the test particle simulation is no exception. The fluctuations present in the error convergence of any Monte Carlo method follow the Central Limit Theorem; thus, they can be reduced simply by averaging together independent ensembles of the same simulation. To reduce these fluctuations and better illustrate the expected convergence of the method, 512 ensembles of the test particle Monte Carlo simulation are collected to obtain the results in Figure 5.9.

As noted in Chapter III, a QMC simulation has the potential of achieving a near-linear error convergence rate that is superior to the $\mathcal{O}\left(N^{-1/2}\right)$ convergence of the test particle Monte Carlo simulation. The QMC method tries to attain this improved convergence by replacing the usual pseudo-random sequence used in Monte Carlo with a sequence that is more uniformly distributed throughout the domain being sampled (*i.e.* a low-discrepancy sequence). A better distribution of sample points is thus expected to yield a better approximation to the integral being sampled. An obvious first attempt to develop a QMC particle simulation for the free molecular conductance probability would be to use a one dimensional low-discrepancy sequence to generate the random variates $u \in \mathscr{U}(0,1)$ instead of the PRN generator. The sequence produced by the PRN generator and the one dimensional low-discrepancy sequence are both uniformly distributed in the unit interval. However, with respect to the star-discrepancy of the two sequences, the low-discrepancy sequence is expected to be significantly more uniform.[29] To test this initial approach for developing a QMC particle simulation, the van der Corput sequence in base 3 (see Appendix A)

[29]Because elements of a low-discrepancy sequence are more uniformly distributed than a random sequence of equivalent length, they are sometimes referred to as "sub-random" sequences in the literature (see [146, 147]).

is used in place of the PRN generator in the test particle simulation. Despite the best intentions of the design, Figure 5.9 clearly demonstrates that the test particle simulation using the van der Corput sequence in base 3 does not converge to the correct solution, for the $L = 2$ duct geometry.

The dismal performance of this first attempt at a QMC particle simulation leads to the obvious question, "What went wrong?" The problem in the test particle simulation can be traced back to the definition of the random variates $u \in \mathscr{U}(0,1)$ needed to produce the sample trajectories. A random variate $u \in \mathscr{U}(0,1)$, by definition, is uniformly distributed in the unit interval and each random variate is independent of all the other variates generated. It is the latter part of this definition that causes the convergence problems for the low-discrepancy sequences. While a good PRN generator is designed such that each number produced appears independent of any previous number generated, the design of the low-discrepancy sequence is exactly the opposite. As noted by Press and Teukolsky in [146], the elements of a low-discrepancy sequence effectively "know" the location of all the other elements in the sequence, and each new element is added so as to "maximally avoid" all the previous elements. It is by virtue of this design that the low-discrepancy sequences are able to achieve a more even distribution of points than a random sequence; yet at the same time, the elements of each dimension of a low-discrepancy sequence are highly dependent on each other. Therefore, the elements of a one dimensional low-discrepancy sequence can not accurately represent a random variate $u \in \mathscr{U}(0,1)$ because they fail to satisfy the necessary independence condition.

The test particle simulation using the van der Corput sequence is physically inconsistent because the interdependence of the low-discrepancy sequence elements prevents them from accurately representing random variates. To better understand

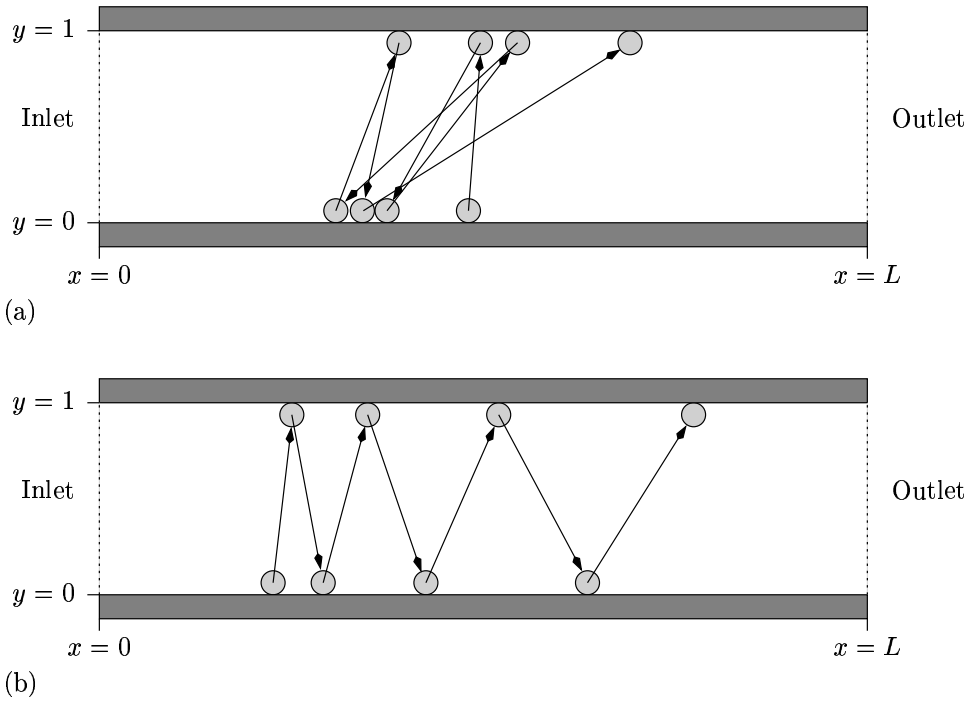

Figure 5.10: Illustration of the non-physical molecular movement within the duct associated with a one dimensional low-discrepancy sequence: (a) using the base 2 van der Corput sequence $(\chi_2(17), \ldots, \chi_2(23))$; and (b) using the base 31 van der Corput sequence $(\chi_{31}(17), \ldots, \chi_{31}(23))$.

this loss of physical accuracy, actual particle trajectories are displayed in Figure 5.10 for the test particle simulation using the van der Corput sequences in bases 2 and 31. The particles simulated using the van der Corput sequence in base 2 alternate directions after each move, which causes the trajectory in Figure 5.10(a) to bounce back and forth around a central location in the duct. This is a direct consequence of the construction of the sequence because any pair of consecutive elements x_i, x_{i+1} of the van der Corput sequence in base 2 is always split into opposite halves of the unit interval. That is, $x_i < \frac{1}{2}$ and $x_{i+1} \geq \frac{1}{2}$, or vice-versa. Since the trajectory locations generated by (5.42) move the particle toward the outlet when the random variate $u > \frac{1}{2}$ and toward the inlet when $u < \frac{1}{2}$, the particle behavior in Figure 5.10(a) is produced.

The particle behavior is much different for the test particle simulation using the van der Corput sequence in base 31. Figure 5.10(b) illustrates that the test particle trajectory generated by part of this sequence marches toward the outlet with increasingly longer reflections. As before, the pathological behavior in the particle trajectory is due to the specific construction details of the van der Corput sequence. Consider the following block of consecutive elements of the van der Corput sequence in base 31: $\{x_n, x_{n+1}, \ldots, x_{n+30}\}$, with $n \equiv 0 \pmod{31}$. For each of these blocks, the first element x_n is found somewhere within the interval $[0, \frac{1}{31})$, and the remaining elements $x_{n+1}, \ldots, x_{n+30}$ are exactly $\frac{1}{31}$ greater than the previous element (*i.e.* $x_{i+1} = x_i + \frac{1}{31}$ for $i = n+1, \ldots, n+30$). As a consequence, the trajectory locations generated by (5.42) move the particle with more oblique reflections toward the outlet when produced by the second half of the block $\{x_{n+16}, \ldots, x_{n+30}\}$, as shown in Figure 5.10(b). Conversely, the trajectory locations produced by the first half of the block $\{x_n, \ldots, x_{n+14}\}$ move the particle with more acute reflections toward the inlet. The particle behavior demonstrated for the test particle simulation using the van der Corput sequence in base 31 is similar to that found for any van der Corput sequence in relatively large base. Using the van der Corput sequence in base 2 and 31, the resulting particle behavior in both cases is a grossly inaccurate representation of the true physical collision process occurring at the wall.

A better approach to designing a QMC particle simulation would be to replace the one dimensional low-discrepancy sequence tested with a multi-dimensional low-discrepancy sequence. For a sufficiently long sequence[30], the coordinates of each multi-dimensional sequence element appear as uniformly distributed, independent variates in the unit interval; albeit with a much more uniform distribution than

[30]A more thorough discussion on the length of a low-discrepancy sequence necessary for the coordinates of the sequence to appear independent, or uncorrelated, is given in Section 6.4.

a random sequence. Given that the dimensions of a low-discrepancy sequence are independent, each coordinate of an $(e+1)$-dimensional sequence element can be used to generate a distinct trajectory location without producing the non-physical behavior demonstrated in Figure 5.10. More specifically, if $\mathbf{x}_n = (x_{1,n}, \ldots, x_{e+1,n}) \in \overline{I}^{e+1}$ denotes the n^{th} element of an $(e+1)$-dimensional low-discrepancy sequence, then the trajectory locations in $T^{(n)}$ (5.36) are generated by using: $x_{1,n}$ as a random variate for $y_0^{(n)}$, $x_{2,n}$ as a random variate for $z_{1,n}, \ldots$, and so forth. The implementation of the test particle simulation with a multi-dimensional low-discrepancy sequence requires some additional consideration because the number of independent variates needed for each sample trajectory $e+1$ is not constant. This is somewhat problematic for most low-discrepancy sequences because they are most efficiently generated from the previous sequence element, which requires all the sequence dimensions to be generated, even if not used by the simulation. Two options are available to handle the varying number of particle moves per trajectory encountered in the test particle simulation: (i) generate a low-discrepancy sequence using a sufficiently large number of dimensions to accommodate all possible trajectory lengths to be simulated; or (ii) adopt a less efficient technique for generating the low-discrepancy sequence that allows each dimension to be generated as needed. While option (i) uses a much more efficient technique for generating the sequence, it tends to be extremely wasteful as the average number of trajectory moves is many times smaller than the number of moves in the longest trajectory. Because the additional cost of generating the extra dimensions of the sequence as needed is less than generating the entire sequence using option (i), option (ii) is selected for the Halton sequence.

The test particle simulation using the Halton sequence is used to approximate the conductance probability of a free molecular duct with a length to height ratio $L = 2$.

Figure 5.9 demonstrates that the relative error of the test particle simulation using the Halton sequence appears to converge, unlike the test particle simulation using the one dimensional van der Corput sequence. In this case, using a multi-dimensional low-discrepancy sequence has eliminated the previous problem of generating independent random variates for the test particle simulation. The test particle simulation using the Halton sequence yields a lower error than the traditional test particle Monte Carlo method when the sample size $N > 256$. Furthermore, the test particle simulation using the Halton sequence is almost an order of magnitude more accurate than Monte Carlo, when the sample size $N > 10^5$. While substituting the PRN generator with the Halton sequence improves the performance, the error convergence rate of the test particle simulation using the Halton sequence is nowhere near the theoretical near-linear error convergence rate of a QMC method.

Unfortunately, the test particle simulation in its current form can not use the Koksma-Hlawka inequality (3.5) to bound the simulation error because the integrand of the problem is not of bounded variation in the sense of Hardy and Krause. Without the Koksma-Hlawka inequality, however, there is no guarantee that the test particle simulation will have near-linear asymptotic convergence. Any multi-dimensional function with a discontinuity not-aligned with the principle axes is not of bounded variation in the sense of Hardy and Krause, as noted in Chapter III. The presence of the YES/NO decisions in the generation of the sample trajectories produce a tremendous amount of discontinuities in the integral representation[31] of the test particle simulation; and almost all of the discontinuities are not aligned with

[31]The formal integral representation of the test particle simulation is not given here; however, an approximate integral form is given as the summation in (5.38). A YES/NO decision is made whenever the test particle simulation checks if a trajectory location z_i is still within the interior of the duct $[0, L]$. This process then appears in the integrand as the discontinuous indicator function used to score the trajectory $S_{mc}(T)$ in (5.38).

the principle axes. When a function contains these pathological discontinuities, it is common that the QMC approximation of its integral will fail to achieve near-linear convergence, and will offer only a slightly better convergence rate than the Monte Carlo approximation (see [115, 117, 120]). To obtain a QMC particle simulation with a higher error convergence rate, one is motivated to find an alternate formulation of the problem that eliminates the YES/NO decisions in the simulation that lead to discontinuities. One such formulation can be found using a classic variance reduction technique presented in the following subsection.

5.5.2 Absorption Weighted Monte Carlo Method

The absorption weighted Monte Carlo (AWMC) method is a variance reduction technique that is nearly as old as the Monte Carlo method itself. One of the earliest applications for the AWMC method is the simulation of radiation transport for shielding applications (see [166] for an example). In particular, the amount of radiation that escapes containment is important to know for safety considerations. The process of the radiation shield absorbing an energetic particle is probabilistic in nature; the goal of a good shield is to absorb as many energetic particles as possible. If the vast majority of the energetic particles are absorbed by the shield, then only a relatively small fraction of the total particles simulated by the Monte Carlo method make a contribution to the escape estimate. This, unfortunately, makes obtaining an accurate escape estimate very costly. To achieve a more efficient simulation, the AWMC method essentially prevents the stochastic absorption process from occurring along the trajectory of the energetic particle. For example, if there is a 90% chance an energetic particle will be absorbed by the shield along a given segment of its trajectory, then the simulated particle weight is reduced by a factor of 10.

Because the particle weight is reduced by the probability of not being absorbed, the AWMC method remains physically consistent with the original problem. All the trajectories generated in the AWMC method are then able to contribute to the escape estimate. However, the contribution of each absorption weighted trajectory is much less than the trajectories that escape the original Monte Carlo simulation because of the reductions in the simulated particle weight.

The equivalent probabilistic absorbing process in the test particle simulation corresponds to the random escape of particles through the inlet and outlet of the duct. Thus, the absorption weighted (AW) formulation of the test particle simulation prevents the simulated particles from leaving the duct interior, thereby eliminating the YES/NO decisions present in the original simulation. To maintain a physically consistent particle trajectory, the weight of the simulated particles must be reduced each time they collide with the wall. Specifically, the particle weight is reduced by the probability that the particle would escape from its given location during its next move, if it were allowed to do so. Fortunately, this probability can be easily calculated for any position within the duct using the outlet escape probability distribution $h(x)$ (5.24) developed for the Nyström method in Section 5.4. Due to the symmetry present in the duct geometry and the distribution of the trajectory angles $f_\theta(\theta)$ (5.5), the inlet escape probability $g(x) = h(L - x)$. Since the simulated particles of the AW method are not allowed to escape the duct, let alone the outlet, a new trajectory score must also be used to estimate the conductance probability. The new trajectory score is then equal to the sum, over all trajectory moves, of the probability that a particle escapes the duct through the outlet during a given move, which can be determined from the outlet escape probability distribution $h(x)$ (5.24) as well. The resulting simulation using the AW formulation is physically consistent with the

original test particle simulation with the following two important improvements: (i) the variance of the simulation is lower; and (ii) the YES/NO decisions in the original formulation have been eliminated.

To proceed more formally, for a duct with a length to height ratio L define

$$\overline{T}^{(n)} = \left\{ z_1^{(n)}, z_2^{(n)}, \ldots, z_s^{(n)} \right\}, \tag{5.43}$$

as the n^{th} particle trajectory of the AW simulation, where $z_1^{(n)}, \ldots, z_s^{(n)} \in [0, L]$ are the points along the interior walls where the particle collides. There are two key differences between the AW trajectory $\overline{T}^{(n)}$ (5.43) and the test particle Monte Carlo trajectory $T^{(n)}$ (5.36): (i) the first collision with the wall interior $z_1^{(n)}$ of the AW trajectory is calculated without considering where the particle first intersects the inlet plane; and (ii) the number of particle moves s is fixed for all AW trajectories. In general, the number of particle moves s per trajectory is taken to be sufficiently large so as to ensure that the weight of the simulated particle at the end of the trajectory is negligible. Each trajectory is given a score $S_{awmc}(\overline{T}^{(n)})$ to determine its contribution to the estimate of the conductance probability. Specifically,

$$S_{awmc}(\overline{T}^{(n)}) = \rho(L) + \sum_{i=1}^{s} w_i h(z_i^{(n)}), \tag{5.44}$$

where $\rho(L)$ (5.15) is the probability a particle escapes the duct directly from the inlet with no wall collisions, w_i is the weight of the simulated particle at the i^{th} location of the trajectory, and $h(z)$ (5.24) is the probability a particle directly escapes through the outlet from z on the next move. The initial weight w_1 of the simulated particle at the first collision with the wall z_1 must exclude the fraction of particles that would normally escape the duct directly; that is, $w_1 = 1 - \rho(L)$. Each subsequent particle weight w_i (for $2 \leq i \leq s$) must exclude the fraction of particles that would normally

escape the duct (through both the inlet and outlet) from the location z_{i-1}. Hence,

$$w_i = w_{i-1}\big(1 - h(L - z_{i-1}) - h(z_{i-1})\big).$$

The AW simulation then yields the following approximation to the conductance probability Ψ given by

$$\Psi = \frac{1}{N}\sum_{n=1}^{N} S_{awmc}(\overline{T}^{(n)}), \tag{5.45}$$

where N is the total number of sample trajectories simulated.

The process of scoring the trajectories for the AW method is illustrated in Figure 5.11 in an effort to better understand $S_{awmc}(\overline{T})$ in (5.44). At the start of the i^{th} trajectory move, the simulated particle is located on the duct wall at $z_i \in [0, L]$ with a particle weight w_i (for $i = 1, \ldots, s$). For each trajectory location z_i, the probability $g(z_i) = h(L - z_i)$ of a particle escaping the duct from z_i through the inlet directly (*i.e.* with no other wall collisions) is calculated. Note that this probability is equivalent to the fraction of particles that would escape from z_i during the next move of the test particle simulation. Also calculated is the probability $h(z_i)$ (5.24) of a particle escaping the duct from z_i directly through the outlet. Because the simulated particles in the AW method are not allowed to escape the interior of the duct, the fraction of particles that would normally escape from z_i in the test particle simulation must be accounted for by reducing the particle weight; *i.e.* $w_{i+1} = w_i(1 - h(L - z_i) - h(z_i))$. For the same reason, one must add the fraction of the simulated particle $w_i h(z_i)$ that would normally escape the outlet to the trajectory score in (5.44) for the AW method. This process of eliminating the particle fraction that would normally escape the duct (in the test particle simulation) from the simulated particle weight is illustrated in Figure 5.11(a).

After the inlet $g(z_i) = h(L - z_i)$ and outlet $h(z_i)$ escape probabilities have been

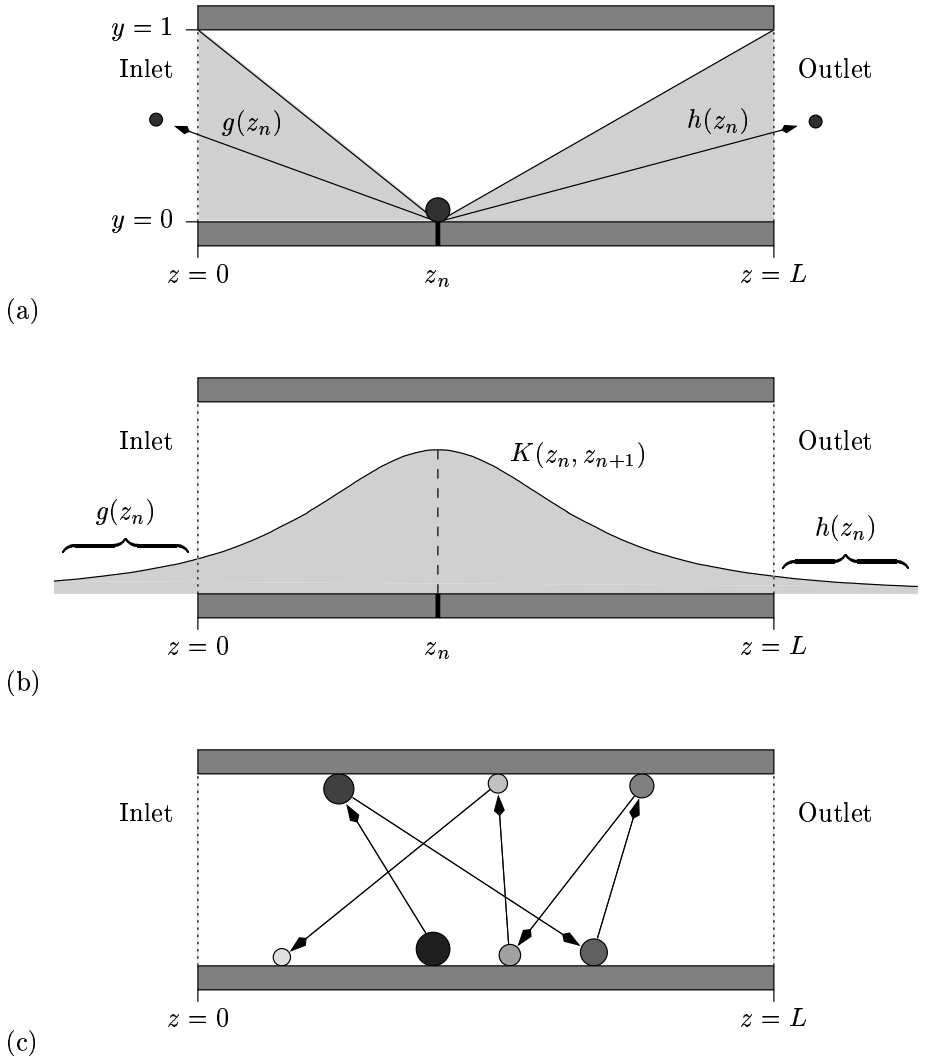

Figure 5.11: Generating the sample trajectory for the absorption weighted (AW) method: (a) removing the fraction of particles that escape the duct from the simulated particle weight; (b) finding the next trajectory location from the probability distribution function $K(z_n, z_{n+1})$ and excluding the particles that escape the inlet $g(z_n) = h(L - z_n)$ and the outlet $h(z_n)$; and (c) the particle trajectory is updated repeatedly until the weight of the particle remaining in the duct is reduced to a negligible level.

found, the next wall collision z_{i+1} can be generated for the AW sample trajectory. In the test particle simulation, the new trajectory locations are generated from the transition probability distribution $K(\Delta z)$ (5.41). In order to remain physically consistent with the problem, the AW method uses the same distribution $K(\Delta z)$ and inverse cumulative distribution $\mathcal{K}^{-1}(u)$ (5.42) as the test particle simulation. The generation of the new trajectory location from $\mathcal{K}^{-1}(u)$, however, must be slightly modified to prevent the simulated particles from escaping the interior of the duct. In particular, the new trajectory location z_{i+1}, generated in (5.42), is less than zero (*i.e.* escaped through the inlet) whenever the random variate $u \in \mathscr{U}(0,1)$ is in the interval $[0, g(z_i))$. Similarly, the new trajectory location z_{i+1} is greater than L (*i.e.* escaped through the outlet) whenever the random variate $u \in (1-h(z_i), 1)$. A physically consistent trajectory is thus produced for the AW method by simply restricting the uniform variate u to the interval $[g(z_i), 1-h(z_i)]$. That is, the trajectory moves $z_2, \ldots, z_s$ of the AW method are generated using $\mathcal{K}^{-1}(u)$ (5.42) with a random variate $u \in \mathscr{U}(g(z_i), 1-h(z_i))$, as shown in Figure 5.11(b). The process of generating each new location of the AW trajectory is repeated until the final trajectory location z_s (see Figure 5.11(c)). At this point along the AW trajectory, the particle weight w_s remaining in the duct should be negligibly small so as not to affect the approximation of the conductance probability in (5.45).

The procedure for determining the first wall collision z_1 must be stated in order to completely define the generation of the sample AW trajectories $\overline{T}$ in (5.43). As noted previously, the first wall collision z_1 does not rely on the initial location y_0 at which the particle crosses the inlet plane. Instead, the initial probability distribution function $b(z)$ (5.25) developed for the Nyström method is used to generate z_1 directly. The initial probability distribution function $b(z)$ gives the probability

a particle collides in the infinitesimal neighborhood of z on the wall plane directly from the inlet without undergoing any other wall collisions. Therefore, the inverse cumulative distribution function $\mathcal{B}^{-1}(u)$ of $b(z)$ can be used to generate the location of the first wall collision z_1; specifically,

$$z_1 = \mathcal{B}^{-1}(u) = \frac{u(2-u)}{1-u}, \tag{5.46}$$

where u is a uniformly distributed random variate.

As with the other particle moves generated for the AW trajectory, some care must be used when selecting the interval over which the random variates in (5.46) are distributed. The initial probability distribution function $b(z)$ gives the probability that the particle first intersects the wall plane at any location $z_1 \in (0, \infty)$, including outside the duct. The fraction of particles that first intersect the wall plane at $z_1 > L$ is simply the direct escape probability $\rho(L)$ given in (5.15). To prevent the simulated particle from escaping the duct on the first move, the uniformly distributed random variate u in (5.46) is therefore restricted to the interval $[0, 1-\rho(L)]$; *i.e.* $u \in \mathscr{U}(0, 1-\rho(L))$. As a consequence of preventing any simulated particles from escaping on the first move, the fraction of particles $\rho(L)$ that would normally escape the test particle simulation directly must also be added to the AW trajectory score in (5.44). The AW simulation is now completely defined by these two basic steps for producing a sample trajectory in (5.43): (i) generating the first intersection with the wall plane z_1 using (5.46) with a random variate $u \in \mathscr{U}(0, 1-\rho(L))$; and (ii) generating the subsequent moves z_i (for $i = 2, \ldots, s$) between the interior walls of the duct using (5.42) with a random variate $u \in \mathscr{U}(g(z_{i-1}), 1-h(z_{i-1}))$.

No method is prescribed for generating any of the random variates $u \in \mathscr{U}(0,1)$ needed in the preceding discussion of the AW particle simulation. The absorption

weighted Monte Carlo (AWMC) method is obtained when the random variates $u \in \mathscr{U}(0,1)$ are produced from a pseudo-random number (PRN) generator scaled to the appropriate interval. Figure 5.9 demonstrates that the relative error in the conductance probability Ψ does indeed converge for the AWMC particle simulation of a free molecular duct with a length to height ratio $L = 2$. In this case, there are 45 particle moves for each sample trajectory in the AWMC simulation, leaving at the end of the trajectory an average particle fraction of approximately 10^{-9} in the duct. The truncation error in the AWMC approximation caused by the leftover particle fraction is negligibly small compared to the overall simulation error. To reduce the fluctuations present in the error convergence, that occur in almost all Monte Carlo applications, 32 ensembles of the AWMC method are collected to obtain the results in Figure 5.9. The expected relative error of the AWMC method is nearly 4 times smaller than the test particle Monte Carlo method for the same number of samples N. The lower simulation error for the AWMC method is a direct consequence of the lower variance in the AW trajectory scores. In fact, the variance σ^2_{awmc} of the trajectory scores in AWMC method is nearly 16 times smaller than the variance σ^2_{mc} of the trajectory scores in the test particle simulation. Note that, as a result of the Central Limit Theorem [47], the factor by which the relative error decreases in the AWMC method is approximately equal to the square root of the variance ratio $\sigma^2_{mc}/\sigma^2_{awmc}$. The convergence rate of the AWMC simulation is $\mathcal{O}\left(N^{-1/2}\right)$ (where N is the number of sample trajectories generated), which, because it is still a Monte Carlo method, is expected.

Figure 5.12 illustrates the distribution of the trajectory scores for the AWMC particle simulation of different duct length to height ratios L, including the $L = 2$ case simulated in Figure 5.9. For reference, the distribution of the trajectory scores

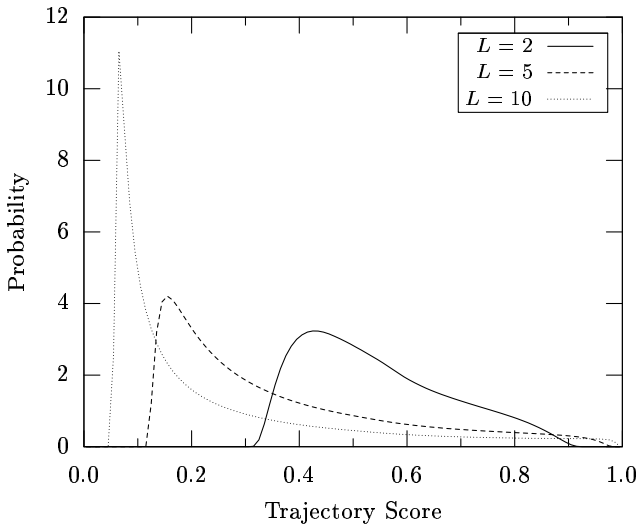

Figure 5.12: Distribution of the trajectory scores for the AWMC particle simulation of free molecular duct flows (for $L = 2$, 5 and 10).

for the test particle Monte Carlo method (5.37) is represented by two delta functions at zero $\delta(x)$ and one $\delta(x-1)$. The weight of the delta function $\delta(x-1)$ is simply equal to the conductance probability Ψ, and the weight of the delta function $\delta(x)$ must necessarily be equal to $(1-\Psi)$. As such, the weight of the delta function at zero increases as the duct becomes narrower (*i.e.* L increases). Moreover, the variance of the trajectory scores in the test particle Monte Carlo simulation is given by $\sigma_{mc}^2 = \Psi(1-\Psi)$. By comparison, most of the trajectory scores for the $L = 2$ duct geometry are relatively close to the mean trajectory score ($\Psi \approx 0.542$), thereby resulting in a much lower variance than the test particle Monte Carlo simulation. Note in Figure 5.12 that as L increases, the distributions of trajectory scores for the AWMC method become more skewed toward a score zero with the majority of the scores found in a narrower band. Thus, the trajectory score distributions for the AWMC method appear to approach the distribution of the test particle Monte Carlo method as $L \to \infty$, which indicates the amount of variance reduction attained by the AWMC method is likely to decrease in this limit.

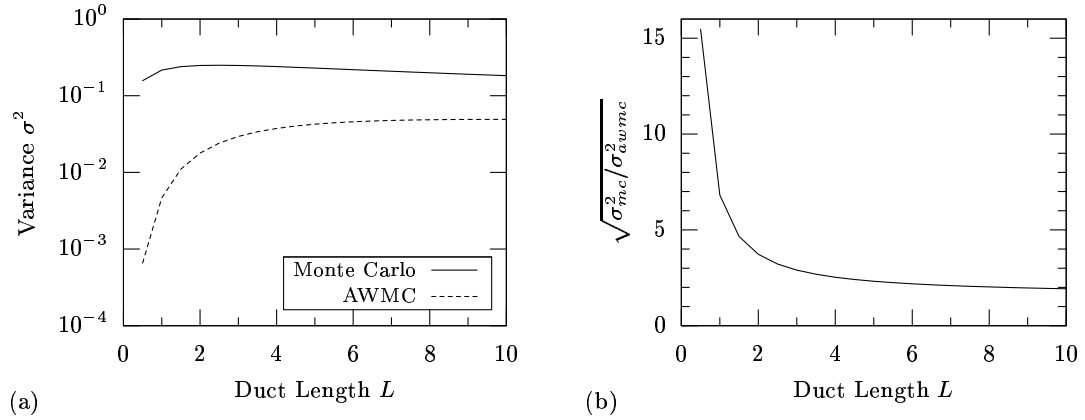

Figure 5.13: Comparison of the test particle Monte Carlo simulation and the AWMC simulation (for $0.5 \leq L \leq 10$): (a) variance σ^2 in the trajectory scores; and (b) the approximate reduction factor for the relative error of the AWMC particle simulation.

Figure 5.13(a) further supports the observation that the amount of variance reduction achieved by the AWMC method decreases as the duct length to height ratio increases. The variance of the AWMC method is nearly 240 times smaller than that of the test particle Monte Carlo method when $L = 0.5$. As the duct becomes narrower, however, this reduction factor monotonically decreases until the variance of the AWMC method is only 3.7 times smaller than the test particle Monte Carlo method when $L = 10$. The corresponding reduction factor between the relative error of the two methods is given in Figure 5.13(b). The AWMC method, in any application, is most effective when the rate of absorption is high. Consequently, the reduction factor between the relative error of the two methods increases significantly as $L \to 0$ because the probability of a simulated particle escaping these wider ducts also increases. The relative error reduction factor monotonically decreases at slower rate in the other limit as the duct length to height ratio L increases; specifically, $\sqrt{\sigma^2_{mc}/\sigma^2_{awmc}}$ decreases from 3 to 1.9 over the range $3 \leq L \leq 10$, as shown in Figure 5.13(b).

It is important to note that the mere fact that the AWMC method has a lower relative error than the test particle Monte Carlo simulation does not make it the superior method in terms of computational cost. While the relative error is lower in the AWMC method for the same number of sample trajectories, the computational cost of generating each trajectory is much larger. The increased computational cost of the AW trajectories is attributed to two factors: (i) the number of particle moves in the AW trajectory is typically much larger than the average number of particle moves in the test particle trajectory; and (ii) there is additional computational overhead in the calculation of the individual particle moves of the AW trajectory because the uniformly distributed random variates must be rescaled for each move. The former is more significant, and is responsible for most of the increase in computational cost of the AW trajectories for the duct geometries considered in this investigation. Up until this point in the investigation, the actual number of particle moves needed for the AW trajectories has not been stated explicitly. The number of interior particle moves determines the average particle fraction that remains in the duct at the end of the AW trajectory. Since this particle fraction is simply ignored by the simulation, it does not contribute anything to the trajectory score in (5.44). It thereby produces a truncation error in the estimate of the conductance probability in (5.45). As noted in Section 5.4, the L_2 norm of the transition probability kernel $\|K\|_2$ is approximately equal[32] to the probability of a particle remaining within the duct during its next move. Therefore, the average leftover particle fraction for an AW trajectory with s interior moves can be estimated by $(1-\rho(L))\|K\|_2^s$.

In Figure 5.14(a), the number of particle moves s per trajectory is calculated for the AWMC simulation of the duct geometries $0.5 \leq L \leq 10$ assuming 3 different

[32]The approximation becomes more accurate as the number of particle moves increases (see Figure 5.6).

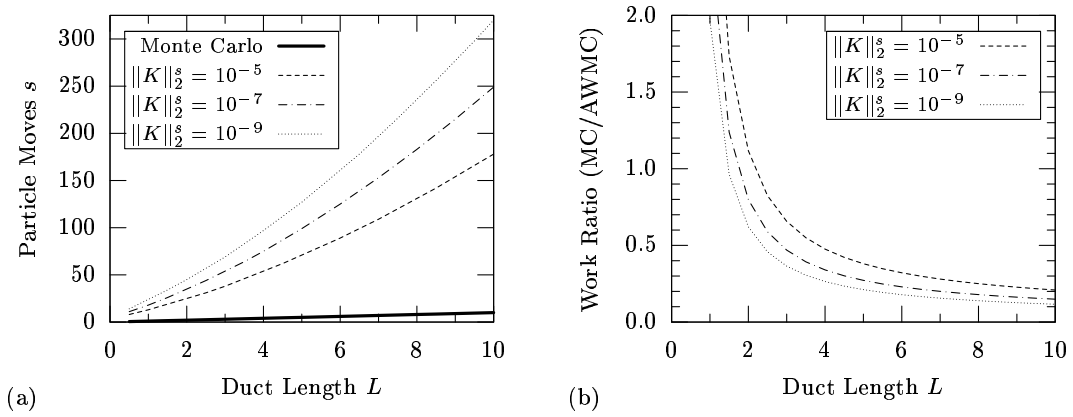

Figure 5.14: Comparison of the test particle Monte Carlo simulation and the AWMC simulation (for $0.5 \leq L \leq 10$): (a) average number of particle moves per trajectory; and (b) the ratio of the total work (particle moves) needed by the test particle Monte Carlo simulation over the AWMC simulation to achieve the same error level.

values of $\|K\|_2^s$. Also found is the average number of particle moves in the test particle Monte Carlo simulation. The narrower duct geometries (*i.e.* $L > 5$) in the AWMC simulation require nearly two orders of magnitude more particle moves than the test particle Monte Carlo simulation. However, the relative error in the AWMC simulation is less than 3 times smaller than the test particle simulation for these duct geometries (see Figure 5.13(b)), seeming to indicate that the AWMC simulation is slower overall. To obtain a better estimate of the computation time required for the two methods, the following work ratio is considered,

$$\text{Work Ratio} = \frac{N_{mc} s_{mc}}{N_{awmc} s_{awmc}} = \frac{\sigma_{mc}^2 s_{mc}}{\sigma_{awmc}^2 s_{awmc}}. \tag{5.47}$$

Here N_{mc} and N_{awmc} are the expected number of sample trajectories needed by the two methods (test particle Monte Carlo and AWMC respectively) to achieve the same error level, and s_{mc} and s_{awmc} are the expected number of particle moves per trajectory needed by the two methods. Note that the work ratio only considers the total number of trajectory moves required by the two simulations, and does not include any cost differences that may arise in the actual generation of the moves.

If the work ratio (5.47) is greater than one for a given duct geometry, the AWMC particle simulation is computationally more efficient than the test particle Monte Carlo simulation; otherwise, the test particle simulation is the better method. Using the average number of particle moves given in Figure 5.14(a) and the variance of the two methods in Figure 5.13, the work ratio (5.47) is calculated in Figure 5.14(b) for the duct geometries $0.5 \leq L \leq 10$. The AWMC particle simulation, as anticipated, is only computationally superior for the widest free molecular ducts simulated (*i.e.* $L \leq 1.5$). This is verified, at least in part, from the actual timing results for the $L = 2$ duct geometry presented in Section 6.1.

5.5.3 Quasi-Monte Carlo Method

The first two attempts at developing a QMC method from the test particle simulation in Section 5.5.1 failed to produce a simulation with a near-linear error convergence rate. The test particle simulation using a one dimensional low-discrepancy sequence is not even physically consistent, as the independent random variates needed for the simulation are not properly represented by the elements of the sequence. The test particle simulation using a multi-dimensional low-discrepancy sequence converges to the correct solution; even at a rate slightly faster than the Monte Carlo method. This rate is not even close to being as fast the theoretical near-linear convergence demonstrated for some QMC applications, however. The problem with the test particle simulation using a multi-dimensional low-discrepancy sequence, as noted earlier, is caused by the discontinuities produced by the YES/NO decisions in the trajectory scoring function S_{mc} (5.37).

Sarkar and Prasad [153] encounter a similar problem with the presence of discontinuities in their development of a QMC simulation for a model radiation transport

problem. They adopt in [153] an alternate formulation of the radiation problem based on the classic variance reduction technique of absorption weighting. The resulting radiation transport simulation no longer requires any discontinuous YES/NO decisions to be made during the generation of a sample. The error convergence rate in [153] approaches the theoretical near-linear limit when a multi-dimensional low-discrepancy sequence is used to generate the random variates in their absorption weighted simulation. Since the absorption weighted particle simulation similarly eliminates the discontinuous YES/NO decisions in the generation of the sample trajectory, the approach of Sarkar and Prasad [153] is adopted here.

The third and final attempt at developing a QMC particle simulation of free molecular flow is based on the absorption weighted particle simulation developed in Section 5.5.2. The QMC particle simulation uses the same form for the sample trajectories $\overline{T}^{(n)}$ (5.43), and the same trajectory score $S_{awmc}(\overline{T}^{(n)})$ (5.44) as the AWMC method in order to estimate the conductance probability Ψ using (5.45). The only difference is that, instead of a PRN generator, the QMC particle simulation uses a low-discrepancy sequence to generate the random variates needed to produce the sample trajectory in (5.43). More specifically, if $\mathbf{x}_n = (x_{1,n}, \ldots, x_{s,n}) \in \overline{I}^s$ denotes the n^{th} element of an s-dimensional low-discrepancy sequence, then the trajectory locations in $\overline{T}^{(n)}$ (5.43) are generated by using $x_{1,n}$ as a random variate for $z_1^{(n)}$, $x_{2,n}$ as a random variate for $z_{2,n}, \ldots$, and so forth. It is important to remember that the random variates generated by the low-discrepancy sequence must also be appropriately scaled to the correct interval in order to prevent the simulated particles from escaping the duct. The simulation procedure is now completely defined, and it is this final version presented here that is referred to as the QMC particle simulation throughout this investigation.

In order to test this new QMC particle simulation, a multi-dimensional low-discrepancy Halton sequence is used to generate the random variates needed for the sample trajectory. The number of particle moves in each sample trajectory and thus, the dimension of the Halton sequence, is taken to be 45, which leaves an average particle fraction of approximately 10^{-9} in the duct at the end of the trajectory. Figure 5.9 demonstrates that the relative error in the conductance probability Ψ converges for the QMC particle simulation of a free molecular duct with a length to height ratio $L = 2$. More interesting is the fact that the error convergence rate of the QMC particle simulation is nearly linear with the number of sample trajectories N. The accuracy of the QMC particle simulation is clearly superior to the traditional test particle Monte Carlo method, and after $N = 10^8$ sample trajectories the relative error of the QMC particle simulation is nearly three orders of magnitude smaller (see Figure 5.9). Because of the near-linear error convergence rate, the cost (in terms of the number of sample trajectories[33]) is dramatically less for the QMC particle simulation. For example, the QMC particle simulation in Figure 5.9 is able to achieve a relative error of 10^{-3} and 10^{-4} using fewer than $4{\cdot}10^3$ and $5{\cdot}10^4$ sample trajectories, respectively. In stark contrast, the Monte Carlo test particle simulation requires more than $5 \cdot 10^5$ and $5 \cdot 10^7$ sample trajectories, respectively, to achieve the same expected relative error levels. Most importantly, Figure 5.9 demonstrates that it is possible to develop a free molecular QMC particle simulation with near-linear error convergence.

Consider the following two dimensional projections of the trajectory scoring functions $S_{mc}(T)$ (5.37) and $S_{awmc}(T)$ (5.44), in order to better understand why the QMC particle simulation based on the absorption weighted trajectory yields better convergence. Let $T(x_1, x_2)$ represent the test particle trajectory generated when

[33]The actual cost difference in terms of computational time is given in Section 6.3.

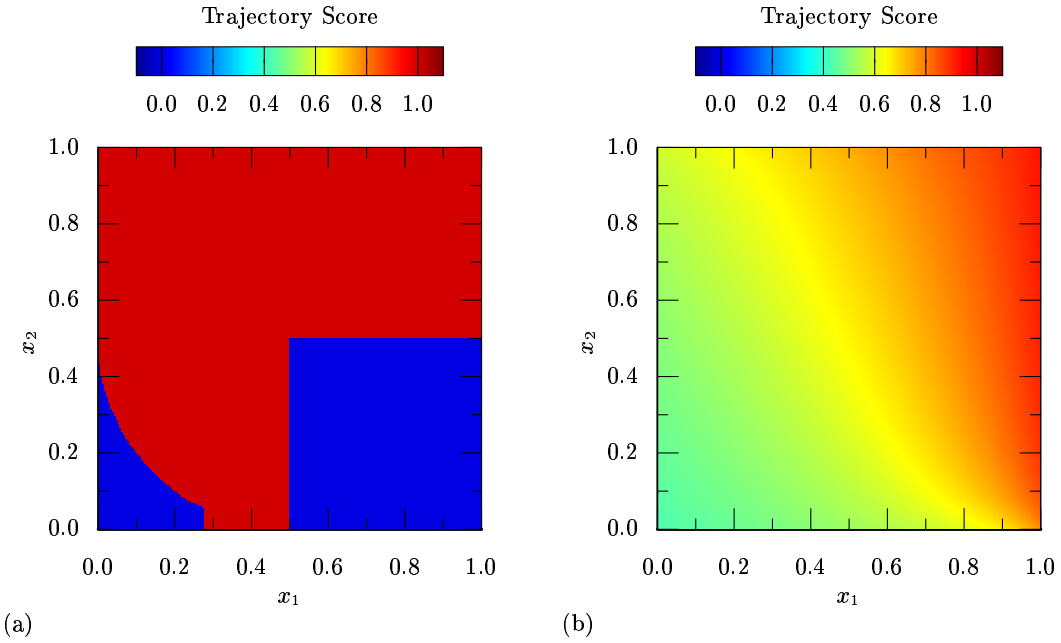

Figure 5.15: Two dimensional projections of the trajectory score functions: (a) for the test particle simulation $S_{mc}(T(x_1, x_2))$; and (b) for the absorption weighted simulation $S_{awmc}(T(x_1, x_2))$.

$x_1, x_2 \in [0, 1)$ are taken as the random variates used to produce the first two wall collisions z_1 and z_2 in (5.36); and all the other locations $y_0, z_3, \ldots, z_e$ are generated assuming the random variates are equal to one. Similarly, let $\overline{T}(x_1, x_2)$ represent the absorption weighted trajectory generated when $x_1, x_2 \in [0, 1)$ are taken as the coordinates of the low-discrepancy sequence used to produce the first two wall collisions z_1 and z_2 in (5.43); and all the other locations $z_3, \ldots, z_s$ are generated assuming the other sequence coordinates are equal to one.[34] The distribution of possible trajectory scores of the test particle simulation $S_{mc}(T(x_1, x_2))$ is given in Figure 5.15(a); and the distribution of possible trajectory scores of the test particle simulation $S_{awmc}(\overline{T}(x_1, x_2))$ is given in Figure 5.15(b). Because of the manner in which these projections of the trajectory scoring functions are defined, the variation

[34]Note that each coordinate of the low-discrepancy (or pseudo-random) sequence $x_i \in [0, 1)$ for $1 \leq i \leq s$ must be properly scaled from the unit interval to a smaller interval in order to prevent the particles from escaping the AW simulation.

in the sense of Vitali (3.43) of $S_{mc}(T(x_1, x_2))$ and $S_{awmc}(\overline{T}(x_1, x_2))$ directly affects the overall variation in the sense of Hardy and Krause (3.45) of their respective scoring functions. In particular, the variation in the sense of Vitali of $S_{mc}(T(x_1, x_2))$ and $S_{awmc}(\overline{T}(x_1, x_2))$ represent one term in the summation (3.45) for $V_{HK}(S_{mc})$ and $V_{HK}(S_{awmc})$; and thereby serve as a lower bound for the variation in the sense of Hardy and Krause.

Note that there exists a discontinuity that is not aligned with the principle axes (x_1, x_2) present in the two dimensional projection of the test particle trajectory scoring function illustrated in Figure 5.15(a). Any function with a discontinuity not aligned with the principle axes is not bounded in the sense of Vitali (see Section 3.3, and also [117, 120]). Consequently, the two dimensional projection $S_{mc}(T(x_1, x_2))$ of the test particle scoring function is not of bounded variation in the sense of Vitali, implying that the overall variation of the scoring function S_{mc} in (5.37) is not bounded in the sense of Hardy and Krause. Without a finite bound on the variation of S_{mc} in the sense of Hardy and Krause, the Koksma-Hlawka inequality cannot be used to bound the error convergence of the test particle simulation given in (5.38). Thus, there is no theoretical means to establish that the test particle simulation will achieve near-linear error convergence when used as a QMC method. In contrast, the two dimensional projection of the AW trajectory scoring function illustrated in Figure 5.15(b) is continuous with a bounded variation in the sense of Vitali approximately equal to 0.1. The bounded and continuous distribution of trajectory scores in Figures 5.12 and 5.15(b) are a good indication that the scoring function S_{awmc} (5.44) for the AW trajectories is likely to be of bounded variation in the sense of Hardy and Krause. However, this graphical evidence by no means constitutes a rigorous proof. It is possible, using the basic properties of multi-dimensional variation reviewed by Owen

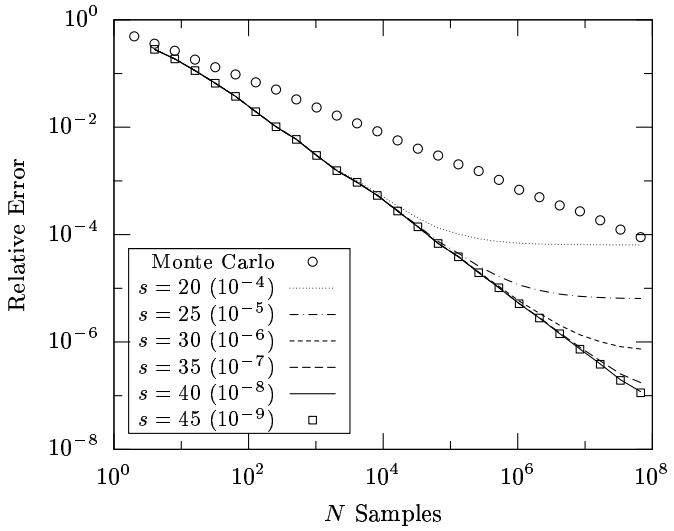

Figure 5.16: Effect of the number of particle moves s used per sample trajectory in the QMC simulation on the overall error of the method. The parenthetical quantities are an estimate $\|K\|_2^s$ of the particle fraction remaining in the duct at the end of the trajectory.

in [136], to prove that the scoring function S_{awmc} is indeed of bounded variation in the sense of Hardy and Krause. Therefore, the error of QMC particle simulation developed in this investigation is bounded by the Koksma-Hlawka inequality; and as such, the method achieves a near-linear error convergence rate in the theoretical limit as the number of sample trajectories $N \to \infty$.

Up until this point, the effect of the number of particle moves in the AW trajectory on the truncation error in (5.44) and the overall simulation error in (5.45) have only been briefly addressed. The QMC particle simulation serves as an excellent test case to illustrate this effect in greater detail because of the great accuracy achieved by the method. In particular, the QMC particle simulation is performed using between 20 and 45 particle moves per trajectory for a free molecular duct with a length to height ratio $L = 2$. The convergence of the relative error of these QMC particle simulations is given in Figure 5.16. Note that the parenthetical quantity in the legend of Figure 5.16 is the estimate $\|K\|_2^s$ of the leftover particle fraction that remains in the duct

at the end of the sample trajectory. It is clear that the overall accuracy of the QMC particle simulation improves as the number of particle moves s per sample trajectory increases. When the number of particle moves s is less than 40, the simulation error in Figure 5.16 appears to stop converging after reaching an error level approximately equal to $\|K\|_2^s$. This indicates that the truncation error caused by the leftover particle fraction dominates the overall error when the error convergence begins to level off.

In practice, the exact number of particle moves required for a given QMC particle simulation depends on the desired accuracy of the application. For example, there is no discernible difference between the error of the QMC simulation using $s = 45$ ($\|K\|_2^s \approx 10^{-9}$) particle moves per trajectory and using $s = 40$ ($\|K\|_2^s \approx 10^{-8}$) particle moves per trajectory. Both simulations have the same error convergence and are able to achieve a relative error of 10^{-7} for the $L = 2$ duct geometry. There is, however, some deviation in the error convergence from the $s \geq 40$ simulations, occurring at an error level of 10^{-6}, if the number of particle moves is reduced further to $s = 35$ ($\|K\|_2^s \approx 10^{-7}$). These results suggest the following rule of thumb, which is adopted for all the remaining QMC particle simulations tested in this investigation: the number of particle moves s per sample trajectory is selected such that $\|K\|_2^s$ is at least 10 times smaller than the smallest simulation error anticipated and/or desired.

CHAPTER VI

RESULTS FOR FREE MOLECULAR DUCT FLOW

The QMC particle simulation developed in Section 5.5 for the free molecular conductance probability is tested for several duct geometries with a length to height ratio L in the range $0.5 \leq L \leq 10$. In particular, the QMC particle simulation is implemented with four of the low-discrepancy sequences presented in Section 4.3: the Weyl-Richtmyer sequence, the Halton sequence, the Faure sequence, and the Niederreiter sequence in base 2. Because its construction and implementation is nearly the same as the Niederreiter sequence in base 2, the Sobol' sequence is not tested here. The only substantial difference between the two methods is that the Niederreiter sequence in base 2 has a slightly smaller constant in the asymptotic discrepancy bound, which is why it is selected here over the Sobol' sequence.[1] The error convergence is found for the QMC particle simulation for each duct length to height ratio L tested and compared to the traditional test particle Monte Carlo method, as well as the absorption weighted Monte Carlo method discussed in Section

[1]Both the Niederreiter sequence in base in base 2 and the Sobol' sequence are constructed from polynomials over the finite field $\mathbb{F}_2$. The Niederreiter sequence uses irreducible polynomials, while the Sobol' uses primitive polynomials. The constant in the asymptotic discrepancy bound grows with the degree of the polynomials $\mathbb{F}_2[x]$ used in the construction. Every primitive polynomial is irreducible; however, the converse is not always true. Thus, there are more irreducible polynomials of low degree available than primitive polynomials resulting in a lower asymptotic error bound for the Niederreiter in base 2 sequence.

5.5.

The QMC particle simulations yield significant performance gains over the traditional Monte Carlo methods in terms of both computational cost and accuracy for most of the duct geometries tested. However, based on the error convergence data presented here, the performance gains tend to diminish as the duct length to height ratio L increases. This observed performance loss of the QMC particle simulation is attributed to the increase in the problem dimension as the duct narrows, which then requires an increase in the dimension of the low-discrepancy sequence. The impact of the dimensionality of the low-discrepancy sequence on the accuracy of the QMC particle simulations is discussed in terms of a non-physical correlation that is present between the problem dimensions. The extent and magnitude of this correlation is calculated here. Given that the performance of the QMC particle simulation suffers as the problem dimension grows, a hybrid QMC/Monte Carlo method is developed to reduce the effective dimension simulated by the QMC method. While the hybrid QMC/Monte Carlo method does not actually improve the accuracy of the original QMC particle simulation, it is able to achieve the same accuracy much faster.

A brief outline of the chapter organization is as follows. In Section 6.1, the conductance probability is found for the free molecular duct with a length to height ratio $L = 2$ using the QMC and Monte Carlo particle simulations. The error convergence of the methods is compared, and a significant reduction in the total computation time is demonstrated for the QMC particle methods. In Section 6.2, the conductance probability is calculated for free molecular flow through more narrow duct geometries ($L = 5$ and $L = 10$); and similarly, the error convergence of the particle methods is compared. In Section 6.3, a more detailed length study of 20 duct geometries in the range $0.5 \leq L \leq 10$ are tested by the particle methods in an effort

to better understand the loss of performance observed in the QMC simulations. The key difference between these results and the previous two sections is that the performance of the particle methods is quantified by the constants of a power law model fit to the error convergence data. This allows for the performance of the particle methods to be compared for all duct geometries on a single plot. In Section 6.3, the impact of correlation between the dimensions of the low-discrepancy sequences is studied. The magnitude and extent of the correlation is found to increase with the number of dimensions in the low-discrepancy sequence, and is the likely cause of the performance loss of the QMC particle simulation. Finally, in Section 6.5, a hybrid QMC/MC method is introduced which decreases the effective dimension of the problem simulated by the QMC portion of the method. The resulting hybrid QMC/MC method shows that is is possible to reduce the computation time of the original QMC particle simulation without affecting the accuracy of the method.

6.1 The $L = 2$ Case

The performance of the QMC particle simulation developed in this investigation for free molecular duct flow is first tested for a short duct with a length to height ratio $L = 2$. In the earliest stages of this investigation, the $L = 2$ duct geometry was the first to be successfully simulated with QMC method using the low-discrepancy Halton sequence; the results of which are presented by McNenly and Boyd in [110]. The initial selection of the $L = 2$ duct geometry was not motivated by any specific physical concerns. Rather, it was selected to ease the algorithm debugging commonly associated with developing the new simulation. While its initial selection was perhaps by happenstance, with the benefit of hindsight the $L = 2$ duct geometry actually serves as an excellent starting point for the performance discussion of the QMC

particle simulation. Specifically, the QMC particle simulation developed here for the $L = 2$ case, using a moderate number of low-discrepancy sequence dimensions ($35 \leq s \leq 50$), clearly demonstrates the superior error convergence rate of the method compared to the traditional Monte Carlo techniques. Before proceeding to the actual performance of the QMC particle simulation for the $L = 2$ geometry, it is necessary to address some of the specific simulation details that are common to all of the results presented in this chapter. In particular, any discussion of the error convergence of a particle simulation requires the following two questions to be answered: first, how should the simulation error be measured; and second, how many samples should be simulated?

As to the first question, "how should the simulation error be measured?", the relative error[2] in the particle simulation is found for the free molecular conductance probability Ψ of the given duct geometry. Recall from the central limit theorem that the particle simulations using pseudo-random sequences do not have a deterministic bound on the simulation error. Thus, for a fixed number of samples, there is a probabilistic confidence interval on the error in traditional Monte Carlo test particle method and the absorption weighted Monte Carlo (AWMC) method. Since each independent[3] simulation using the pseudo-random sequences can produce a range of approximate solutions, ensembles of these simulations are collected to estimate the average, or expected, simulation error. For the traditional Monte Carlo test particle method, 512 independent ensembles of the simulation are collected for each case presented in this chapter. For the AWMC method, which has a lower variance than

[2]The relative error is the difference between the simulation solution and the exact solution normalized by the exact solution. Here the "exact" solution is taken from the more accurate Nyström method which is shown in Section 5.4 to have a stable relative error less than 10^{-12} for the duct lengths under consideration.

[3]That is, the pseudo-random number generator for an independent simulation uses a distinct seed value to initialize the sequence.

the traditional Monte Carlo method (see Figure 5.13), only 32 independent ensembles of the simulation are collected for each case presented in this chapter. In contrast, it is possible to obtain a deterministic error bound on the QMC particle simulations from the Koksma-Hlawka inequality (3.5). The relative error found in this section for the $L = 2$ duct geometry is calculated using a single ensemble of the QMC particle simulations. However, it is important to note that, in most cases, the Koksma-Hlawka inequality does not provide a tight upper bound on the integration error of the QMC particle simulation. Hence, to obtain a better estimate of the expected error convergence of the QMC particle simulation, it is common to collect ensembles of the simulation based on independent subsequences of the same low-discrepancy sequence. While not used in this section, ensembles of the QMC particle simulation are collected for the general performance study in the next section (please refer to Figures 6.7-6.10) to provide a more accurate representation of the error convergence rate.

As to the second question, "how many samples should be simulated?", the Central Limit Theorem provides a probabilistic bound on the simulation error that monotonically decreases with the sample size for the traditional test particle Monte Carlo method and the AWMC method. Thus, the actual number of samples needed for these Monte Carlo simulations can be estimated in advanced if the variance of the simulation is known. Unlike the Monte Carlo simulation, the lack of a tight error bound provided by the Koksma-Hlawka inequality makes it difficult to determine from theory alone the number of samples needed for the QMC particle simulation in practice. In particular, Morokoff and Caflisch note in [116] that the dominant term $N^{-1}(\log N)^s$ appearing in the discrepancy bound for most of low-discrepancy sequences does not actually decrease until $N > e^s$. Hence, for the dimension of the

low-discrepancy sequences used in the QMC particle simulations, the error bound from the Koksma-Hlawka inequality is not yet decreasing for the sample sizes N under consideration. However, as previously demonstrated in the development of the method (see Figure 5.16), the QMC particle simulation still converges for these sample sizes in practice, albeit with no theoretical assurance that the error will monotonically decrease. In contrast, for the traditional test particle Monte Carlo method and the AWMC method, the Central Limit Theorem provides at least a probabilistic bound on the simulation error that monotonically decreases with the sample size. As a result, one may wonder if checking the QMC simulation error at a specific number of samples N is a representative estimate of the overall convergence rate compared to, *e.g.* $N-1$, $N+1$, $N+2047$, or any other value of N. More importantly, from an implementation and performance standpoint, one may ask the following. Are there specific values of the sample size N (or equivalently the sequence length) which are known *a priori* to yield a lower than average integration error or sequence discrepancy? For some low-discrepancy sequences, the answer is, in fact, yes.

As an example, consider the van der Corput sequence in base 2 that is constructed in Appendix A. Figure 3.8 shows that the star discrepancy of the van der Corput sequence in base 2 clearly achieves the lowest possible value when the sequence length N is a power of 2. This behavior in the convergence of the star discrepancy can be understood from: (i) the concept of the $(t, s)-$sequence introduced by Niederreiter in [125]; and (ii) noting that the van der Corput sequence in the example is a $(0, 1)$ sequence in base 2. From the general definition of the $(t, s)-$sequence in base b given in [127], the distribution of the sequence satisfies a desirable uniformity condition when the sequence length $N = b^{t+1}$. Moreover, this uniformity condition is satisfied for every block of b^{t+1} sequence elements, which suggests that $N = kb^{t+1}$, for $k =$

$1, 2, \ldots$, may yield a better than average approximation from the QMC simulation using a $(t, s) -$ sequence in base b. Another key feature to note is that for each subsequent power of b taken for the sequence length, *i.e.* $N = (b^{t+1}, b^{t+2}, b^{t+3}, \ldots)$, the uniformity condition becomes even stronger.

All the error results presented in this investigation are given on a logarithmic scale to facilitate comparisons between the convergence rates of the different particle simulations. Thus, to achieve a uniform spacing on the logarithmic scale, it is natural to consider sample sizes of $N = (2, 2^2, 2^3, \ldots)$ when performing QMC simulations with the Niederreiter sequence in base 2, or the Sobol' sequence, as both are examples of $(t, s) -$ sequences in base 2. The Faure sequence is another example of a (t, s)-sequence; specifically, it is a $(0, s) -$sequence in base q, where q is the smallest prime greater than or equal to the sequence dimension s. Then, for the same reasons as the base 2 sequence, it appears beneficial to stop the QMC simulation using the Faure sequence for sample sizes of $N = (q, q^2, q^3, \ldots)$. However, the dimension s of the low-discrepancy sequence required for the QMC particle simulations in this investigation is typically too large to simulate all but the first couple of powers $N = (q, q^2, q^3, \ldots)$ because $q \geq s$. Instead, it is more practical to consider sample sizes that have a spacing similar to those tested for other sequences, but with the condition of $N \equiv 0 \pmod{q}$. Since the parameter t equals zero for the Faure sequence, selecting the sample sizes from the set $N = kq$, for $k = 1, 2, \ldots$ ensures that at least the basic uniformity condition is satisfied.

The remaining low-discrepancy sequences considered in this investigation, namely the Weyl-Richtmyer and Halton sequences, can not be classified as $(t, s) -$sequences. This does not, however, preclude the existence of specific sample sizes N (or equivalently sequence lengths) which are known from the construction to have some at-

tributes that may lead to potentially better discrepancy or simulation approximation. For any low-discrepancy sequence, the convergence of the star discrepancy guarantees the existence of a "minimum" set of sequence lengths $\mathcal{M}_{min} = (N_1, N_2, \ldots)$ with the property that $D^*_{N_i} < D^*_N$ for all $N < N_i$ and $i \geq 1$. If actually known, the sequence lengths contained in $\mathcal{M}$ would make favorable points to stop the QMC simulation and check the results. Except in the one dimensional case[4], there is not a constructive method for determining the set $\mathcal{M}_{min}$ for a general Weyl-Richtmyer sequence. Given the analogue between the Weyl-Richtmyer sequence and the method of good lattice points due to Korobov, it may be possible to adopt a similar exhaustive search for sequence lengths that are optimal in some sense. However, as discussed in Section 3.4, such exhaustive searches are computationally intractable for the sequence lengths and dimensions needed in practice for the QMC particle simulation.

Each dimension of the Halton sequence is generated by a distinct van der Corput sequence in a relatively prime base. While the s dimensional Halton sequence used in this investigation is constructed from a series of $(0,1) -$ sequences in bases $p_1, p_2, \ldots, p_s$, where p_i represents the i^{th} smallest prime; the lack of a common base prevents the Halton sequence from being considered a (t,s)-sequence. However, if the sequence length N is selected such that

$$\begin{aligned} N &\equiv 0 \pmod{p_1}, \\ &\vdots \\ N &\equiv 0 \pmod{p_s}, \end{aligned}$$

then, from the definition of the $(0,1) -$ sequence, each dimension of the Halton

[4]By inspecting Theorem 3.3 in [127], the one dimensional Weyl-Richtmyer sequence constructed from an irrational number z achieves a minimum in the extreme discrepancy bound when the sequence length N equals the denominator of the rational convergents of z. Specifically, when N is taken from the set $\mathcal{M}_{min} = (q_1, q_2, \ldots)$, where $r_i = p_i/q_i$ denotes the i^{th} rational convergent to the irrational number z determined from the first i terms of the continued fraction representation of z.

sequence would satisfy the basic uniformity condition of the (t, s)-sequence. These specific values for the sequence length correspond to $N = k \cdot p_1 \cdots p_s$ for $k = 1, 2, \ldots$. Unfortunately, for the number of sequence dimensions s needed for the QMC particle simulations in this investigation, taking the sample size N to be the product of the first s primes is much too large to simulate in practice. As with the general Weyl-Richtmyer sequence, there is no obvious criteria that can be used in practice for the Halton sequence to select sequence lengths N for the QMC simulation that have known beneficial properties, such as a lower than average integration error or sequence discrepancy.

After considering the second question, "how many samples should be simulated?" the relative error of the various particle simulations is found for the following sample sizes, or equivalently, sequence lengths N. In order to ensure the additional uniformity property attributed to the $(t, s) -$ sequences, the relative error of the QMC simulation using the Niederreiter sequence in base 2 is found for the sequence lengths $N = (2, 2^2, \ldots, 2^{26})$. Based on the same reasoning, the relative error of the QMC simulation using the Faure sequence in base 53 is found for the sequence lengths $N = (53, 53 \cdot 2, \ldots, 53 \cdot 2^{20})$. The Halton and Weyl-Richtmyer sequences are not $(t, s) -$ sequences, and there is not a similar uniformity condition to be exploited for the selection of the sequence length N. As such, the relative error in the QMC simulations using these sequences is found for the sequence lengths $N = (2, 2^2, \ldots, 2^{26})$ simply to provide uniform spacing on the logarithmic scale and to be consistent with the other simulations. Unlike the QMC simulations using a (t, s)-sequence, there is no special number of samples for a Monte Carlo simulation that may produce a better than average approximation; besides, of course, the monotonic decrease in the expected error due to the central limit theorem. Thus, for the results in this

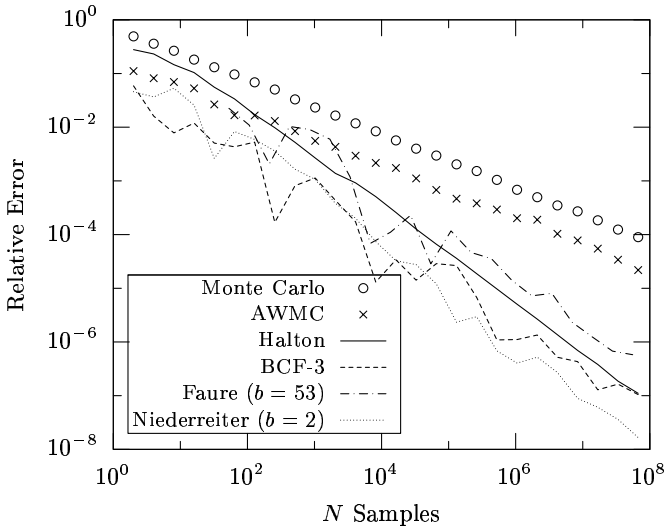

Figure 6.1: Convergence of the relative error for the QMC particle simulation of the conductance probability Ψ (for $L = 2$).

chapter, the traditional Monte Carlo test particle method and the AWMC method are performed for the sample sizes $N = (2, 2^2, \ldots, 2^{26})$ for consistency.

The convergence of the relative error in the conductance probability is given in Figure 6.1 for the Monte Carlo test particle method, the absorption weighted Monte Carlo (AWMC) method, and the QMC particle simulations. The QMC particle simulations are implemented using the four low-discrepancy sequences reviewed in Section 4.3: the Halton sequence, the BCF-3 sequence, the Faure sequence, and the Niederreiter sequence in base 2. In the AWMC and QMC simulations, the number of interior particle moves, and hence the dimension of the sequences is taken to be $s = 50$. After 50 moves, the average fraction of the particle that remains in the duct is approximately $7 \cdot 10^{-11}$, which is two orders of magnitude smaller than the lowest error observed in Figure 6.1 making it an acceptable truncation error. It is apparent from the results in Figure 6.1 that every implementation of the QMC particle simulation demonstrates clear superiority over the Monte Carlo methods in

terms of the error convergence rate.

In order to establish a point of comparison for the QMC simulations, the test particle Monte Carlo method achieves a relative error of $9.0 \cdot 10^{-5}$ at $N = 2^{26}$ samples, and the AWMC method achieves a relative error of $2.2 \cdot 10^{-5}$ for the same number of samples, as shown in Figure 6.1. The error convergence rate of both methods is approximately $\mathcal{O}\left(N^{-1/2}\right)$ as anticipated, since both Monte Carlo methods are implemented using a sequence of pseudo-randomly generated numbers. Recall from Section 5.5 that the process of absorption weighting, *i.e.* the gradual escape of a fraction of the particle during each move, is a common variance reduction technique. As a direct consequence of this lower variance, the AWMC method consistently has a relative error that is more than 4 times smaller than the Monte Carlo simulation given the same number of samples. While the AWMC method offers greater accuracy per sample trajectory than the Monte Carlo method, the computational time required to generate each sample trajectory is also greater because of the additional number of interior particle moves. Initial analysis from Section 5.5 on the amount of computational work required by both Monte Carlo methods to achieve the same error indicates that, for the $L = 2$ duct geometry, the AWMC method requires more work than the Monte Carlo Method (see Figure 5.14). This initial analysis is further validated here by the timing results presented in Figure 6.2, which demonstrate that the AWMC method is slightly slower at reaching the same error as the Monte Carlo method.

The physical interpretation of the free molecular flow is different between the two Monte Carlo simulations. Consequently, each sample trajectory of the AWMC method is more accurate than the test particle Monte Carlo method because of the extra computational work associated with the absorption weighting variance reduc-

tion technique. Since the QMC and AWMC methods are both based on the same physical interpretation of the free molecular duct flow, it is improper to compare the QMC simulation to the less accurate test particle Monte Carlo simulation based on the number of sample trajectories alone. Instead, it is more appropriate to use the results from the AWMC method as the representative Monte Carlo simulation when considering the error convergence of QMC particle simulations relative to the number of sample particles. In Figure 6.1, the QMC particle simulations using the Niederreiter sequence in base 2 and the BCF-3 sequence yield the lowest relative error over all the sample sizes N considered, with the Niederreiter sequence providing consistently the best results at the larger values of N. More specifically, after simulating $N = 2^{26}$ sample trajectories, the QMC particle simulation using the Niederreiter sequence in base 2 is more than 1000 times as accurate as the AWMC method. By comparison, the QMC particle simulations using the Halton and BCF-3 sequences are more than 200 times as accurate as the AWMC method with $N = 2^{26}$ sample trajectories. While the QMC simulation using the BCF-3 sequence yields a more accurate approximation than the AWMC method for all sample sizes tested; it is only when $N > 128$ that the QMC simulation using the Halton sequence consistently yields a better estimate than the AWMC method. When the number of sample trajectories $N > 10^5$, the QMC simulation using the Faure sequence yields the least accurate approximation; however, it is still more than 40 times as accurate as the AWMC method at $N = 2^{26}$.

Except for the QMC particle simulation using the Halton sequence, the error convergence for the other QMC simulations is somewhat erratic, making it difficult to assess the error convergence rate by inspection of Figure 6.1 alone. It is possible to perform a linear least squares fit to the convergence data in order to obtain an

estimation of the convergence rate via the power law exponent in (6.1). However, the power law exponent is found to vary by as much as 5% depending on the number and location of the data points included in the linear least squares fit. The observed sensitivity in the power law exponent to the data in Figure 6.1 is a consequence of using only a single ensemble for each simulation. By collecting more ensembles for the error convergence of each QMC simulation, the expected error convergence is much less erratic, which produces a more reliable estimate of the power law exponent.

To evaluate the simulation performance of the $L = 2$ geometry, the following expected error convergence rates are found from the results in Figure 6.8 using 16 ensembles for each QMC simulation. When 16 ensembles are collected for each QMC simulation, the power law exponent is found to vary by less than 2%. The expected error convergence rate of the QMC simulation using the Niederreiter sequence in base 2 is $\mathcal{O}\left(N^{-1.04}\right)$, which is the fastest of all the simulations for this duct geometry. The QMC simulation using the Halton sequence is the next fastest method, with a near linear average error convergence rate of $\mathcal{O}\left(N^{-0.97}\right)$. While slightly slower, the QMC simulations using the Faure and BCF-3 sequences achieve an average error convergence rates of $\mathcal{O}\left(N^{-0.84}\right)$ and $\mathcal{O}\left(N^{-0.78}\right)$, respectively. However, both are still significantly faster than the Monte Carlo techniques.

While the initial results in Figure 6.1 for the error convergence are encouraging, because of the cost differences associated with the QMC simulation it is important to also consider the simulation error as a function of the total computation time.. In Figure 6.2, the computation time τ is found for each of the particle simulations performed in this section. As noted before, the cost of calculating each trajectory of the AWMC is greater than the savings afforded by the variance reduction, making the traditional Monte Carlo test particle method slightly faster for reaching any expected

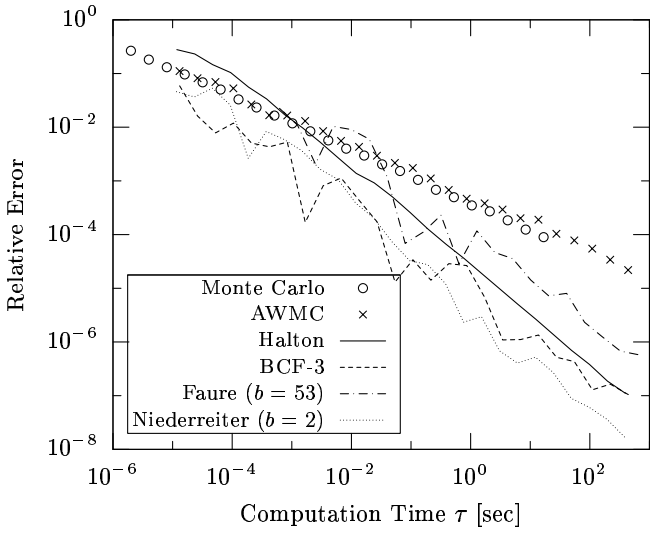

Figure 6.2: Convergence of the relative error with respect to the computation time τ (in seconds) for the conductance probability Ψ (for $L = 2$).

error level. Thus, it is more appropriate to use the results from the test particle method as the representative Monte Carlo simulation when comparing the error convergence of QMC particle simulations relative to the computation time τ. All of the QMC simulations of the $L = 2$ duct geometry yield a higher error convergence rate than the Monte Carlo simulations (*i.e.* the power law exponent $\gamma < -\frac{1}{2}$). Hence, for each QMC simulation, there exists some critical error level E_{crit} to distinguish if the QMC simulation is actually faster than the Monte Carlo test particle method in terms of computation time. For the QMC simulations, the combined effect of the higher convergence rate and higher cost of generating a sample trajectory means that the QMC simulation is expected to achieve any error level below E_{crit} faster than Monte Carlo.

In Figure 6.2, the QMC simulations using the BCF-3 sequence and the Niederreiter sequence in base 2 are the fastest particle methods for reaching any error level below $E_{crit} = 10^{-1}$, which is achieved after only two sample trajectories. The QMC

simulation using the BCF-3 sequence is initially the fastest particle method when the desired accuracy is less than 10^{-3} because the sequence is specifically designed to have a very low power law constant c in the estimated error convergence rate in (6.1). However, the higher error convergence rate and lower sequence generation cost enable the QMC simulation using Niederreiter sequence in base 2 to consistently outperform the BCF-3 sequence when the desired accuracy is greater than 10^{-3}. The QMC simulation using the Halton sequence is the third fastest particle simulation outperforming the Monte Carlo method for any error level below $E_{crit} = 10^{-2}$. It is interesting to note that while the QMC simulation using the Halton sequence possesses a higher error convergence rate than the BCF-3 sequence, the higher cost of generating the Halton sequence results in a slower QMC simulation for almost all the error levels tested here. The QMC simulation using the Faure sequence is consistently the slowest QMC simulation considered because of the large computational cost associated with generating the sequence. In spite of this, the QMC simulation using the Faure sequence is still faster than the test particle Monte Carlo method for reaching any error level below $E_{crit} = 2 \cdot 10^{-3}$.

For the $L = 2$ duct geometry, the QMC simulation using the Niederreiter sequence in base 2 offers perhaps the best combination of accuracy and speed of all the particle simulations. However, in some applications requiring less accuracy, the QMC simulation using the BCF-3 sequence may be slightly faster by virtue of its design. The key result of this section is that it is possible to develop a QMC particle simulation that is significantly faster in terms of computation time than is the traditional test particle Monte Carlo method. In terms of the $L = 2$ duct geometry, the QMC particle simulations developed here also achieve a near-linear error convergence rate in terms of the number of samples N that is superior to the $\mathcal{O}\left(N^{-1/2}\right)$ convergence

of the Monte Carlo methods. Furthermore, the higher error convergence rate enables the QMC simulations to produce approximations that are orders of magnitude more accurate than are obtained with the Monte Carlo methods. This is an especially desirable feature for the particle simulation of low speed rarefied gas flows encountered in many fluidic MEMS applications. While the overall simulation accuracy need not be great, the ability to adequately resolve the very slow average, or bulk, velocity of the gas in the presence of the thermal, or random, speed is critical. For example, consider a fluidic MEMS device with a bulk velocity of 1 m/sec operating in a nitrogen gas environment at standard temperature and pressure.[5] In order to resolve the average velocity to a 10% accuracy level (assuming a 95% confidence interval), traditional DSMC requires more than 15 million independent samples of the flow field to be generated. In contrast, if a general QMC particle method could be developed with linear error convergence, the resulting simulation would only require 4000 samples to achieve the same accuracy.[6]

6.2 The $L = 5$ and $L = 10$ Cases

The results from Section 6.1 for the $L = 2$ duct geometry demonstrate that it is possible to construct a QMC particle simulation with an error convergence rate and computation time superior to the Monte Carlo methods. Based on these encouraging results, it is natural to extend the QMC particle simulation of the free molecular conductance probability to include other duct geometries - especially for narrower ducts that have a larger duct length to height ratio L. This type of duct geometry

[5] At these conditions, the average speed of the nitrogen molecules is $\overline{v} = 455$ m/sec, with a standard deviation $\sigma = 285$ m/sec.

[6] More specifically, this assumes the QMC particle simulation converges as $\mathcal{O}\left(N^{-1}\right)$ with the same implied constant as the Monte Carlo method. In the case of free molecular duct flow, the implied constant for the QMC particle simulations is actually less than that of the Monte Carlo method for nearly all the geometries considered in the next section, as illustrated in Figure 6.9.

commonly occurs for highly non-equilibrium gas flows found in fluidic MEMS, vacuum system designs, and semiconductor manufacturing processes. Here, the error convergence is found for the $L = 5$ and $L = 10$ cases using the Monte Carlo test particle method, the absorption weighted Monte Carlo (AWMC) method, and the QMC particle simulation. As before, a single ensemble of the QMC particle simulation is performed using each of the four low-discrepancy sequences discussed in Section 4.3. Unfortunately, the results for the QMC particle simulations in the narrower duct geometries $L = 5$ and $L = 10$ are not as promising as for the $L = 2$ case. There is a noticeable decrease in the performance of the QMC particle simulations observed as the duct length increases. In fact, only the QMC simulation using the BCF-3 sequence consistently outperforms the AWMC method for the narrower duct geometries presented in this section. A longer duct requires a greater number of interior particle moves; hence, a greater number of dimensions for the low-discrepancy sequence are used in the QMC particle simulation. This problem of dimensionality is well-noted throughout literature for a wide range of QMC applications (see [23, 74, 114, 116, 117, 118, 120, 146, 153, 167]).

In Figure 6.3, the relative error[7] in the conductance probability is found for the particle simulations of the $L = 5$ duct geometry. The relative error is obtained for the same number of samples N as the $L = 2$ case in Section 6.1; adjusting the sample sizes for the QMC simulation with the Faure sequence to reflect the change in the sequence base. In the AWMC and QMC simulations, the number of interior particle moves, and hence the dimension of the low-discrepancy sequences, is taken to be $s = 120$. After 120 moves, the average fraction of the particle that remains

[7]The relative error is the difference between the simulation solution and the exact solution normalized by the exact solution. Here the "exact" solution is taken from the more accurate Nyström method which is shown in Section 5.4 to have a stable relative error less than 10^{-12} for the duct lengths under consideration.

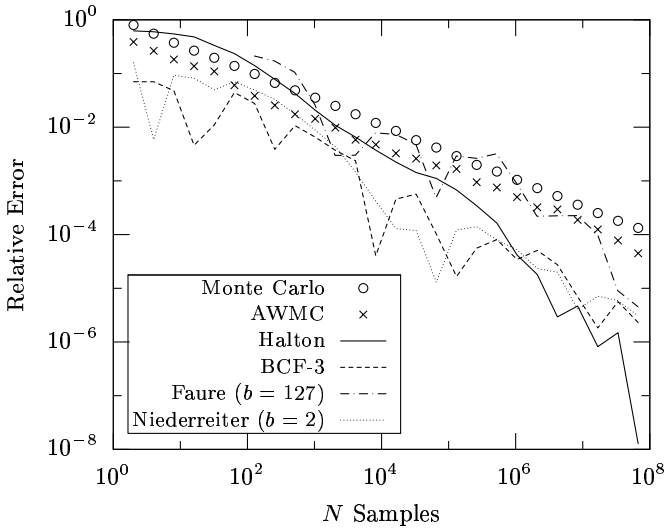

Figure 6.3: Convergence of the relative error for the QMC particle simulation of the conductance probability Ψ (for $L = 5$).

in the duct is approximately $2 \cdot 10^{-9}$, which is an order of magnitude smaller than the lowest error observed in Figure 6.3, making it an acceptable truncation error. As a direct consequence of its lower variance, the AWMC method consistently has a relative error that is nearly 3 times smaller than the Monte Carlo simulation given the same number of samples N. For reference, the lowest relative error achieved by the traditional Monte Carlo test particle method and the AWMC method in this investigation for $N = 2^{26}$ samples is $1.3 \cdot 10^{-4}$ and $4.5 \cdot 10^{-5}$ respectively. Overall, the particle simulations in Figure 6.3 for the $L = 5$ duct geometry have a relative error that is larger than the corresponding simulations for the $L = 2$ case in Section 6.1.

For the $L = 5$ duct geometry, the QMC simulation using the BCF-3 sequence is the only simulation with an error that is consistently smaller than the AWMC method for all sample sizes, as shown in Figure 6.3. The QMC simulation using Niederreiter sequence in base 2 does not have an error consistently smaller than the

AWMC method until the number of samples $N \geq 10^3$. With at least $N = 10^5$ samples, the QMC simulations using the BCF-3 and Niederreiter sequences achieve at least an order of magnitude improvement in the relative error over the AWMC method. The QMC simulation using the Halton sequence eventually reaches the same order of magnitude improvement over the AWMC when the number of samples $N \geq 10^6$. For the longest sequence lengths tested ($N \geq 10^6$), the QMC simulation using the Halton sequence has the lowest error of all the methods; however, the amount and rate by which the error decreases appear to be anomalous. In contrast, the Faure sequences does not appear to offer noticeable improvement over the AWMC method until the number of samples $N \geq 10^7$. Therefore, excluding the results using the Faure sequence, the QMC particle simulations do offer error convergence that is superior to the Monte Carlo methods when the number of samples generated is sufficiently large.

In Figure 6.4, the relative error in the conductance probability is found for the particle simulations of the $L = 10$ duct geometry. The relative error is obtained for the same number of samples N as the $L = 2$ case in Section 6.1. As with the $L = 5$ duct geometry, the sample sizes for the QMC simulation with the Faure sequence are adjusted to reflect the change in the sequence base. In the AWMC and QMC simulations, the number of interior particle moves is taken to be $s = 300$. After 300 moves, the average fraction of the particle that remains in the duct is approximately $2 \cdot 10^{-9}$, which is two orders of magnitude smaller than the lowest error observed in Figure 6.4, making it an acceptable truncation error. The AWMC method, by virtue of its lower variance, consistently has a relative error that is 2.2 times smaller than the Monte Carlo simulation, given the same number of samples N. As a reference point, the lowest relative error achieved by the traditional Monte

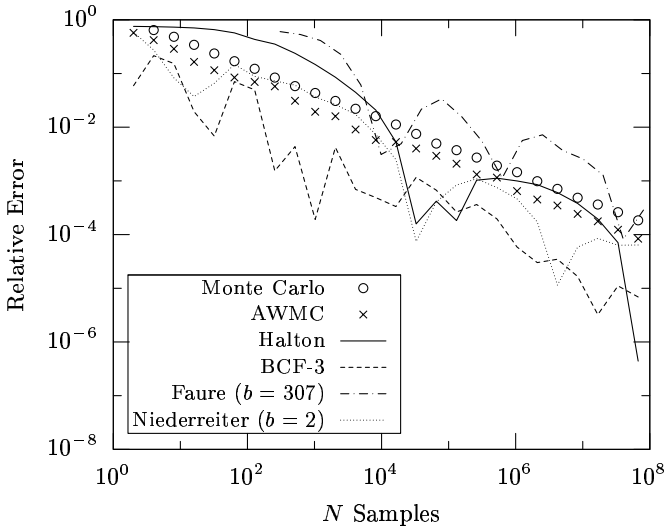

Figure 6.4: Convergence of the relative error for the QMC particle simulation of the conductance probability Ψ (for $L = 10$).

Carlo test particle method and the AWMC method in this investigation, for $N = 2^{26}$ samples, is $1.8 \cdot 10^{-4}$ and $8.4 \cdot 10^{-5}$ respectively. In general, the particle simulations in Figure 6.4 for the $L = 10$ duct geometry have a relative error that is larger than the corresponding simulations for the $L = 5$ case. Moreover, the reduction in the relative error achieved by the absorption weighted technique diminishes as the duct geometry becomes narrower. This observation is further supported by the results in Figure 6.7 in connection with the more rigorous length study presented in the next section.

Regarding the $L = 10$ duct geometry, the QMC simulation using the BCF-3 sequence is again the only simulation with an error that is consistently smaller than the AWMC method for all sample sizes, as shown in Figure 6.4. The QMC simulation using the BCF-3 sequence consistently maintains at least an order of magnitude improvement in the relative error over the AWMC method when the number of samples $N \geq 10^6$. The QMC simulation using Niederreiter sequence in base 2 does not have

an error consistently smaller than the AWMC method until the number of samples $N \geq 10^4$. Even for the larger sample sizes, the QMC simulation using Niederreiter sequence in base 2 rarely offers a significant improvement over the AWMC method. Worse yet are the QMC simulations using the Halton and Faure sequences, which only achieve an error lower than the AWMC method sporadically, with the approximation of the Halton sequence generally better than that of the Faure sequence. Thus, the QMC simulation using the BCF-3 sequence is the only method that achieves an error convergence that is noticeably superior to the two Monte Carlo methods in this particular case.[8] It is important to note that actual behavior of the error convergence for the QMC particle simulation becomes more erratic as the low-discrepancy sequence dimension increases. As a result of this observation, one is motivated to adopt for the QMC simulations the same type of ensemble averaging used to reduce the fluctuations in the error convergence of the Monte Carlo methods. Therefore, 16 ensembles are averaged for each QMC particle simulation in order to illustrate more clearly the impact of the duct geometry on the error convergence, for the more rigorous duct geometry study in Section 6.3.

All low-discrepancy sequences, except for the Weyl-Richtmyer type sequences, suffer from something referred to as *start-up error* (see [18, 19, 48, 116]). Specifically, the start-up error in low-discrepancy sequences refers to the tendency of initial sequence elements to disproportionately cluster near the origin $(0, \ldots, 0)$ of the unit hypercube. When more elements are added to the sequence, the effect of the near-origin clustering lessens as the additional sequence members begin to evenly fill in

[8]Emphasis should be made to note that the BCF-3 sequence is the only method that demonstrates a clear improvement with a single simulation ensemble of the error convergence results. However, using 16 ensembles for the $L = 10$ duct geometry; the expected error at $N = 2^{23}$ samples for the QMC simulations using the Halton and Niederreiter sequences is actually 3.4 times and 5 times more accurate, respectively, than the AWMC method, as shown in Figure 6.7.

the remaining volume of the unit hypercube. In general, an increase in the dimension of low-discrepancy sequence causes a larger start-up error; that is, it takes more sequence elements to average out the initial near-origin clustering of elements. In this investigation, the implementation of the Niederreiter sequence in base 2 attempts to mitigate the start-up error by adopting the "leading zeros" correction proposed by Bratley *et. al.* in [19]. With respect to the Halton and Faure sequences, the start-up error is generally treated by ignoring the first N_{skip} elements of the sequence; *i.e.* the elements disproportionately clustered near the origin [18, 19, 48, 116].

Unfortunately, there is not an established criterion for selecting N_{skip}, and an effective choice depends on the type of sequence, the dimension of the sequence, and the problem type. The implementations of the Halton and Faure sequences in this investigation do not skip any of the initial sequence elements because further study beyond the scope of this investigation is needed to determine suitable choices based on the duct geometry. The presence of start-up error in these sequences is likely to contribute to the slower error convergence shown in Figures 6.3 and 6.4 for the $L = 5$ and $L = 10$ duct geometries. However, in Section 6.3, the use of ensemble averaging to smooth the error convergence of the QMC simulations also serves to diminish the start-up error, since it only appears in the first ensemble. After correcting the startup error for the Halton and Faure sequences, they are expected to yield a relative error similar to the BCF-3 sequence when the number of samples N is small.

There is a marked decrease, in all cases, in the performance of the QMC particle simulations for the duct geometries $L = 5$ and $L = 10$ when compared to the $L = 2$ case in Section 6.1. In addition, the convergence results for a single ensemble of any of the QMC simulations become more erratic as the duct narrows, complicating the estimate of the error convergence rate. In an effort to reduce the fluctuations

found in the error convergence, a more rigorous duct geometry study is presented in Section 6.3 using 16 ensembles of each QMC particle simulation. The results of this more detailed geometry study further support the observation that the QMC simulations become less effective as the duct narrows; that is L increases. There are a greater number of interior particle moves required to produce each sample trajectory of the QMC simulation, when the duct length to height ratio L increases. Consequently, the dimension of the low-discrepancy sequence used by the QMC simulation must also increase. As noted, the performance problems attributed to the increase in the dimension of the low-discrepancy sequence are well-documented throughout literature for many different applications of the QMC method. With respect to the QMC particle simulations developed in this investigation, an increase in low-discrepancy sequence dimension produces an increase in particle behavior that is not physically consistent with the actual problem to be approximated. This connection between the low-discrepancy sequence dimension and simulated particle behavior is explored in greater detail in Section 6.4.

6.3 Duct Geometry Study ($0.5 \leq L \leq 10$)

The goal of this section is to develop a clearer understanding of the performance loss suffered by the QMC particle simulations as the duct length to height ratio L increases. While the preceding results in Sections 6.1 and 6.2 illustrate this performance loss, a more detailed study of the effect of the duct geometry on the QMC performance is presented here. Specifically, using the QMC particle method, the free molecular conductance probability is simulated for 20 different duct geometries in the range of $0.5 \leq L \leq 10$. For reference, the free molecular conductance probability test is also found using the traditional test particle Monte Carlo method, and the

absorption weighted Monte Carlo (AWMC) method. In order to reduce the fluctuations in the QMC results given in Figures 6.1, 6.3, and 6.4, 16 ensembles of each QMC particle simulation are averaged together to produce an expected error convergence of the method. The idea of collecting ensembles of the QMC method to provide a better estimate of its performance is not uncommon, with many examples available in the literature [23, 115, 116, 117, 118]. In this section, five performance metrics are calculated for each particle simulation and duct geometry, which include: (i) the expected relative error after $N = 2^{23}$ sample trajectories; (ii) the error convergence rate; (iii) the single sample error; (iv) the critical error when the QMC particle simulations are faster than the test particle Monte Carlo method; and (v) the computation time speedup of the QMC particle simulations.

In order to estimate the error convergence rate and single sample error of the particle methods, a power law approximation is found for the dependence of the relative error on the sample size N,

$$\frac{|\Psi - \Psi_{part}|}{\Psi} \approx cN^{\gamma}. \tag{6.1}$$

Here Ψ is the exact[9] conductance probability, Ψ_{part} is the conductance probability approximated by the particle simulations, c is the power law constant, and γ is the power law exponent. The power law approximation in (6.1) is found using the standard linear least squares method after performing a logarithmic transformation to the error convergence data, please refer to [147] for a more detailed description. The power law exponent γ is an estimate of the rate, or speed, at which the particle simulation converges; as γ becomes more negative, the simulation is said to converge

[9]Note that the "exact" value of the conductance probability Ψ in (6.1) is obtained from the Nyström method, as described in Section 5.4. The Nyström method is more accurate than the particle methods, and has a stable relative error of at most 10^{-12} for the duct geometries under consideration, as illustrated in Figure 5.7.

faster. The power law constant c is an estimate of the single sample error; when c becomes smaller, each sample trajectory is considered more accurate. In addition, the power law constant also indicates the expected performance during the initial stages of the simulation. That is, after a relatively small number of sample trajectories have been generated, but before the effect of the power law exponent begins to dominate the error convergence.

The error convergence rate of the QMC particle simulations is faster than the Monte Carlo methods, for all the duct geometries tested in this section. This performance gain comes at a price; specifically, the computational cost of generating each sample trajectory of the QMC particle simulation is greater than that of the test particle Monte Carlo method. The increased cost is primarily attributed to the increase in the number of interior particle moves required by the QMC and AWMC simulations. However, as discussed in Section 4.3, there can be an additional cost for generating the actual low-discrepancy sequences in some cases. By virtue of the lower computational cost associated with generating the trajectories, the test particle Monte Carlo method is the fastest particle method when a relatively crude approximation is needed. As the desired accuracy increases, the QMC particle simulation eventually becomes faster than the test particle Monte Carlo method because of the greater error convergence rate, even though there is a higher computational cost per sample. An example of this behavior is found in Figure 6.2 for the QMC simulation of the $L = 2$ duct geometry using the Halton sequence. It is possible to estimate the critical error E_{crit} for each QMC simulation, by combining the power law models of the error convergence with the computation time of the particle simulations. The critical error E_{crit} is the simulation error expected to be reached by both the QMC simulation and the test particle Monte Carlo method in the same amount of com-

putational time. Thus, it serves as the natural transition point between the levels of error at which the QMC simulation is the faster particle method, and vice-versa. For a particular duct geometry, if one needs to calculate the conductance probability with greater accuracy than E_{crit}, then the QMC particle simulation will reach the desired accuracy faster than the test particle Monte Carlo method. Conversely, if one requires only a rough approximation, then the traditional test particle Monte Carlo method is the faster choice. As such, a larger value of E_{crit} implies that the QMC particle simulation is the faster method for a wider range of desired simulation accuracies.

Before proceeding to the performance results of this section, it is necessary to first review the specific simulation details of the geometry study presented here. While averaging additional ensembles of the QMC particle simulation shares the same purpose with the Monte Carlo methods, it is important to note that it does not share the same theoretical underpinnings. The concept of ensemble averaging for Monte Carlo simulations is solidly rooted in the statistical theory of the method. As such, the physical meaning of the ensemble average is well defined in terms of the formal expectation of the error of the method. Unlike the pseudo-random sequences required for the Monte Carlo methods, consecutive elements of the low-discrepancy sequences used by the QMC methods are not designed to appear independent of each other. The elements of a low-discrepancy sequence are distributed as uniformly as possible because, in essence, each new element added to the sequence "knows" the location of all the previous elements, by special construction, avoids placing the new elements too close to any of the previous elements. However, the fact that this "knowledge"exists between the elements means that distinct subsequences are not independent. Consequently, it is not technically correct to interpret the ensemble

average in terms of the formal statistical expectation of the simulation error, when each ensemble of a QMC simulation is produced from a different subsequence of the same low-discrepancy sequence. Instead, one should adopt a more heuristic viewpoint and interpret the results from the ensemble averages of a QMC simulation as an "engineering anticipation" of the performance. Thus, whenever the "expected," or "average" error of the QMC particle simulation is discussed in this investigation, it refers to this heuristic interpretation.

In this section, 16 ensembles of the QMC particle simulation with 2^{23} samples[10] are collected and averaged for each low-discrepancy sequence presented in Section 4.3. A total sequence length $N = 2^{27}$ must be generated for each low-discrepancy sequence in order to produce all 16 ensembles of the QMC particle simulation. The QMC simulation for each ensemble is then performed using distinct subsequences of the low-discrepancy sequence each containing 2^{23} elements. To illustrate the effect of the ensemble averaging, the resulting error convergence is given in Figure 6.5 for the QMC particle simulation of the $L = 10$ duct geometry using the Niederreiter sequence in base 2. The fluctuations in the error convergence for a single QMC simulation appearing in the results in Sections 6.1 and 6.2 are noticeably reduced as the number of simulation ensembles increases. Furthermore, when more simulation ensembles are collected, the power law constant and exponent are less sensitive to the number and location of the data points used to fit the power law model to the error convergence results.

In addition to the 16 ensembles collected for each of the QMC particle simulations, there are 512 simulation ensembles collected for the test particle Monte Carlo method and 32 simulation ensembles collected for the AWMC method. Sim-

[10]Except for the QMC simulations using the Faure sequence in base q which use slightly more samples $N = 2^{23} + r$, where r is the smallest positive integer such that $N \equiv 0 \pmod{q}$.

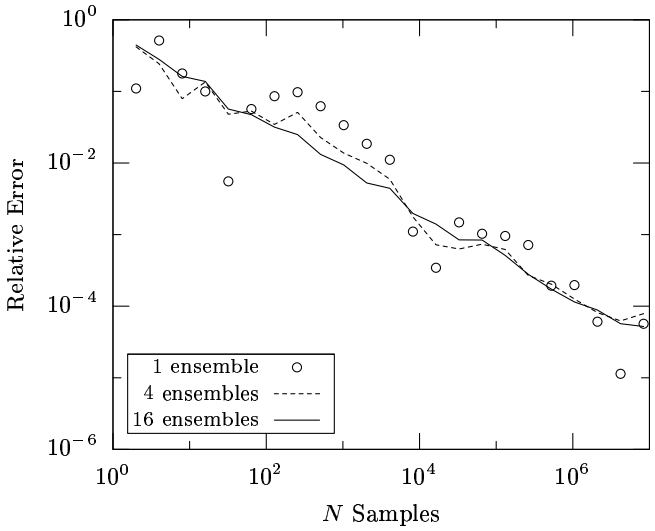

Figure 6.5: Convergence of the relative error after collecting 1, 4, and 16 ensembles for the QMC particle simulation using the Niederreiter sequence in base 2 (for $L = 10$).

ilar to the results in Sections 6.1 and 6.2, the relative error of all the particle simulations (except for the QMC simulation using the Faure sequence) is found for sample sizes $N = (2, 2^2, \ldots, N^{23})$. In the two preceding sections, the relative error for the QMC simulation using the Faure sequence in base q is found for sample sizes $N = (q, 2q, 2^2q, \ldots)$ in an effort to exploit an additional uniformity condition of $(0, s) -$ sequences. However, adopting the same sample sizes for the detailed geometry study allows the length of the Faure sequences to vary by a factor of nearly two for the different duct geometries tested. This is undesirable because it may make the QMC simulations for duct geometries using the longer sequences appear artificially better than others. To avoid this potential problem and preserve the uniformity condition, the QMC simulation using the Faure sequence in base q is found for sample sizes $N = 2^\alpha + r$; where $\alpha = \lfloor \log_2 q \rfloor, \ldots, 23$, and r is the smallest positive integer such that $N = 2^\alpha + r \equiv 0 \pmod{q}$.

Briefly recall from Section 5.4 the integral form of the free molecular conductance

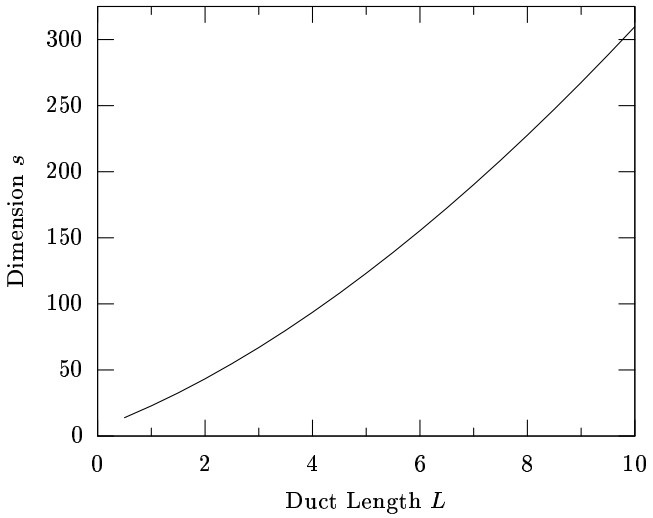

Figure 6.6: The number of interior particle moves (and low-discrepancy sequence dimension s) used for the AWMC and QMC simulations.

problem. The L_2 norm of the integration kernel $\|K\|_2$ in (5.26) can be viewed as the probability a particle inside the duct will not escape through the outlet or inlet during its next move, after undergoing many interior moves.[11] Thus, the probability $\rho(L)$ in (5.15) of a particle directly escaping from the inlet can be multiplied by $\|K\|_2^s$ to obtain a suitable approximation of the expected particle fraction remaining in the AWMC and QMC simulations after s interior particle moves. In this section, the number of interior particle moves (or dimension s of the sequence) needed for the AWMC and QMC particle simulations is selected to satisfy $\|K\|_2^s = 2 \cdot 10^{-9}$. After s moves, the average fraction of the particle that remains in the duct is at least an order of magnitude smaller than the lowest error observed in the detailed geometry study presented in this investigation. It follows that the truncation error of these methods has a negligible impact on the simulation results. In Figure 6.6, this value of the sequence dimension s used in the AWMC and QMC particle simulations of this section is plotted for each of the duct geometries tested.

[11] That is, ignoring the effects of the initial distribution of particles at start-up.

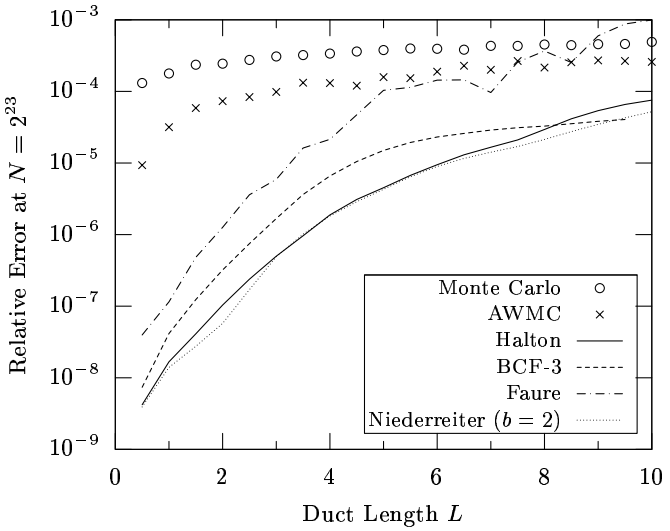

Figure 6.7: The expected relative error of the particle simulations found after generating $N = 2^{23}$ sample trajectories.

In Figure 6.7, the expected relative error of the particle simulations is given for $N = 2^{23}$ sample trajectories. All particle simulations clearly demonstrate an increase in relative error as the duct narrows, *i.e.* L increases, which supports the previous observation made in Section 6.2. The increase in the duct to length ratio L corresponds with an increase in the effective dimension of the problem because, on average, more interior moves are needed before a particle can escape the narrower duct geometry. In general practice, when the dimension of an integral problem increases, the performance, or accuracy, of any numerical method designed to approximate the integral suffers. The accuracy of the traditional test particle Monte Carlo method varies by less than a factor of 5 over the range of duct geometries tested, which makes it the particle simulation most resilient to the negative effects that accompany an increase in the problem dimension. This beneficial feature is common to many different applications of the Monte Carlo method beyond just particle simulations, and explains, in part, the popularity of the method for simulating physical problems with many

dimensions. The accuracy of the AWMC varies over a larger range than the test particle Monte Carlo method. In particular, for the $L = 0.5$ duct geometry, the AWMC is 14 times more accurate than the test particle Monte Carlo method for the same number of sample trajectories; and for the $L = 10$ duct geometry, the AWMC is only 2 times more accurate. It is interesting to note that for the widest ducts, *i.e.* the smallest values of L, the savings due to the variance reduction in the AWMC method is actually large enough for it to be faster than the traditional test particle Monte Carlo method.

Unlike the relatively modest changes in accuracy observed for the Monte Carlo methods, the accuracy of the QMC particle simulations vary by over 4 orders of magnitude depending on the duct geometry. This pronounced impact on the accuracy is attributed to the difficulties encountered when the dimension of low-discrepancy sequences used in the QMC particle simulations is large. The problem of dimensionality appears throughout literature for a wide range of QMC applications [23, 74, 110, 114, 116, 117, 118, 120, 153, 167], and is addressed in greater detail in Section 6.4 for the particle simulations developed here. The most rapid loss of accuracy of the QMC particle simulations occurs within the range $0.5 \leq L \leq 5$. The performance loss is not as significant as the duct becomes narrower ($L > 5$). The QMC simulation using the Niederreiter sequence in base 2 is the most accurate particle method after generating $N = 2^{23}$ samples, for all the duct geometries tested. The QMC simulation using the Halton sequence is a close second, providing nearly the same accuracy as the Niederreiter sequence in base 2 for several of the duct geometries tested. By comparison, the QMC simulation using the BCF-3 sequence is about 2 to 3 times less accurate than the Niederreiter sequence in base 2 for most of the duct geometries considered. As noted, the QMC simulations using the Faure se-

quence are the least accurate of the QMC methods and, in fact, is less accurate than the AWMC method after $N = 2^{23}$ samples and when $L \geq 7.5$. Despite its lackluster performance simulating narrower ducts, the results of the Faure sequence are still at least an order of magnitude more accurate than the AWMC method when $L \leq 2.5$. Ignoring the Faure sequence for a moment, the other QMC particle simulations are all at least 3 orders of magnitude more accurate than the AWMC method for the $L = 0.5$ duct geometry. Despite the rapid loss of accuracy from the $L = 0.5$ case, for the $L = 5$ duct geometry, the QMC simulations are still 37, 35, and 11 times more accurate than the AWMC method using the Niederreiter, Halton, and BCF-3 sequences respectively. It should be noted that the dimension of the low-discrepancy sequence needed for the QMC simulation is 124, for the $L = 5$ case. While there is no strict bound on the number of dimensions that can be simulated with the QMC method, $s = 124$ is within the range generally given for a practical upper limit. The expected error of the QMC simulations using the Niederreiter and BCF-3 sequences is approximately 5 times smaller than the AWMC method, for the $L = 10$ duct geometry when the sequence dimension $s = 310$. This is, perhaps, the largest dimension of the low-discrepancy sequences ever used in practice for an application of the QMC method.[12]

In Figure 6.8, the power law exponent γ, from the linear least squares fit of the error convergence to the power law model in (6.1) is found for each particle simulation and duct geometry tested in this section. As to be anticipated for the Monte Carlo simulations, the power law exponent γ is approximately equal to $-\frac{1}{2}$, regardless of the duct geometry. Note that fluctuations present around the expected convergence rate $\mathcal{O}\left(N^{-1/2}\right)$ are due to the statistical scatter inherent in the method. In contrast

[12]Morokoff [114] uses the low-discrepancy Sobol' sequence in 360 dimensions in a QMC financial simulation; specifically, the valuation of a 30-year bond with monthly coupon payments.

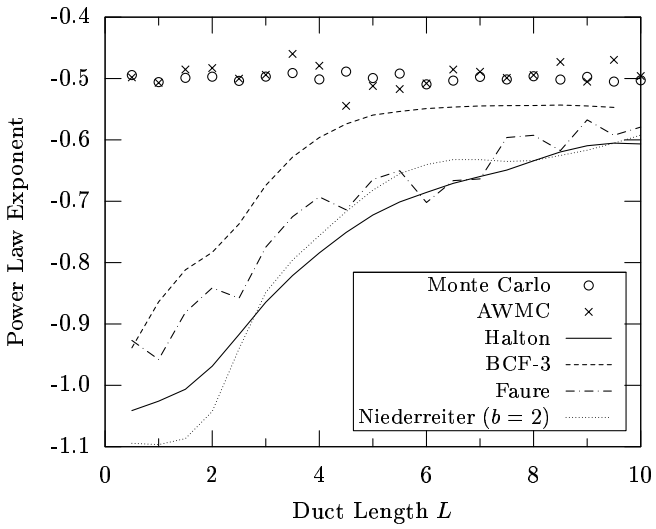

Figure 6.8: The expected error convergence rate of the particle simulations.

with the Monte Carlo methods, the power law exponent for the QMC simulation becomes less negative, *i.e.* the error convergence rate of the methods becomes slower as the duct length to height ratio L increases. The slower error convergence rate of the QMC simulation is the primary cause of the loss of accuracy observed for the method in Figure 6.7. Moreover, the error convergence rate of the QMC simulations slows by the greatest amount within the range $0.5 \leq L \leq 5$, corresponding well to the range of duct geometries where the QMC simulations experience the most rapid loss of accuracy. The error convergence rate is the fastest for the QMC particle simulations using the Niederreiter ($b = 2$) and Halton sequences, as illustrated in Figure 6.8. In particular, the QMC particle simulations using the Halton sequence typically have the fastest error convergence for the narrower duct geometries; specifically, where the duct length to height ratio is $L \geq 2.5$. An important result of this investigation is that the QMC particle simulations using these two low-discrepancy sequences are able to achieve near linear error convergence rates greater than $\mathcal{O}\left(N^{-0.85}\right)$ for the wider duct geometries when $L \leq 3$. The QMC simulation using the Faure sequence tends to

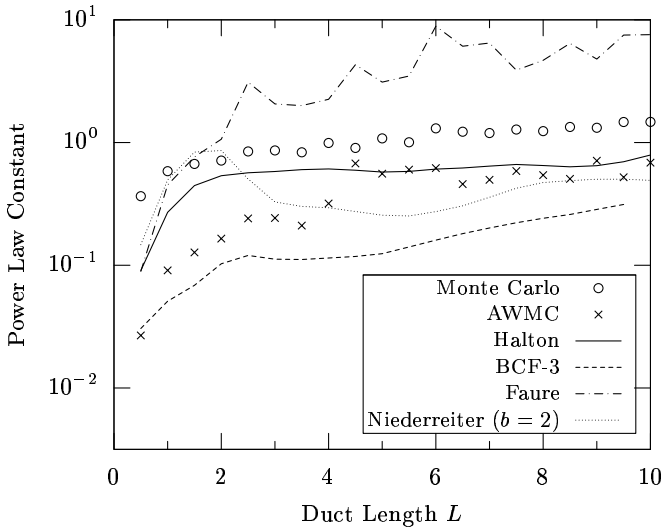

Figure 6.9: The expected single sample error of the particle simulations.

converge at a rate slower than that of the Niederreiter ($b = 2$) and Halton sequences; however, it is still faster than QMC simulation using the BCF-3 sequence.[13] Despite the fact that the error convergence rate of QMC particle simulations slows as L increases, the error convergence rate is still faster than the Monte Carlo methods for even the narrowest duct geometry tested; that is, $L = 10$. More specifically, the error convergence rates using the different low-discrepancy sequences for the $L = 10$ duct geometry are as follows: $\mathcal{O}\left(N^{-0.61}\right)$ using the Halton sequence; $\mathcal{O}\left(N^{-0.59}\right)$ using the Niederreiter sequence in base 2; $\mathcal{O}\left(N^{-0.57}\right)$ using the Faure sequence; and $\mathcal{O}\left(N^{-0.55}\right)$ using the BCF-3 sequence.

In Figure 6.9, the power law constant c (from the linear least squares fit of the error convergence to the power law model in (6.1)) is found for each particle simulation and duct geometry tested in this section. The power law constant is an

[13]While the BCF-3 sequence possesses the slowest convergence of all QMC methods, recall from Figure 4.5 that it is possible to improve the convergence rate using a Weyl-Richtmyer sequence by selecting a different set of irrational numbers. For example, the BCF-5 and Richtmyer sequence converge faster than the BCF-3 sequence in most cases. However, this performance gain comes at the price of having a higher single sample error.

estimate of the expected error present in one sample trajectory. By virtue of the variance reduction technique adopted for the AWMC method, it is not surprising that the AWMC method has a smaller single sample error than the traditional test particle Monte Carlo method. In fact, the difference in the expected single sample error between the two Monte Carlo methods is nearly the same as the difference observed between the relative error of the methods after $N = 2^{23}$ samples (see Figure 6.7). The general trend of the single sample error of the QMC simulations tends to increase as the duct length to height ratio L increases. This contributes to the performance loss observed for the QMC particle simulations. However, except for the Faure sequence implementation, the single sample error only varies by a factor between 5 and 10 for all the QMC simulations, with most of the variation occurring when $L \leq 2$. Thus, the impact on the QMC performance of this increase in the single sample error is not as significant as the decrease in the error convergence rate observed previously in Figure 6.8. In Figure 6.9, for all but the $L = 0.5$ duct geometry, the QMC simulations using the BCF-3 sequence have the smallest single sample error of all the particle simulations. Recall from Section 4.3 that the BCF-3 sequence is selected as the representative Weyl-Richtmyer sequence in this investigation because it consistently has the smallest single sample error of the Weyl-Richtmyer sequences tested in Figure 4.5. Thus, the BCF-3 sequence not only offers the lowest single sample error of the Weyl-Richtmyer sequences tested, but of all the low-discrepancy sequences tested here as well. The single sample error of the QMC simulations using the Niederreiter ($b = 2$) and Halton sequences is approximately the same order as the Monte Carlo methods, with the Niederreiter ($b = 2$) sequence achieving a lower value when $L \geq 2.5$. In particular, the QMC simulation using the BCF-3 sequence has a single sample error that is 1.7 to 2.2 times smaller than the Niederreiter ($b = 2$)

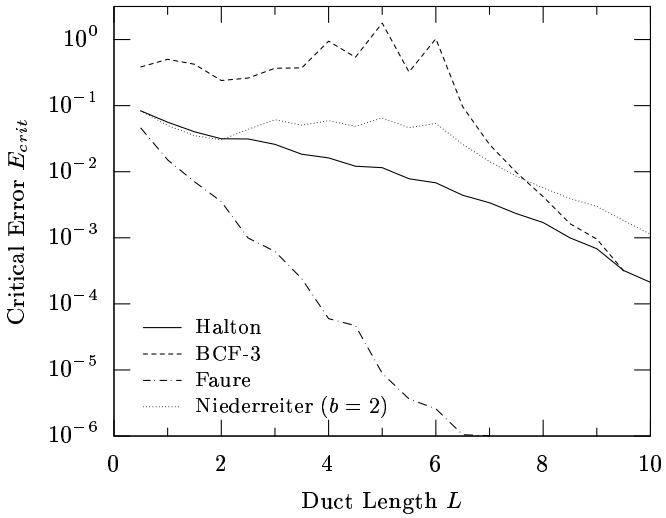

Figure 6.10: The critical error E_{crit} of the QMC particle simulations. If the desired simulation error is less than E_{crit}, the QMC simulation is faster than the test particle Monte Carlo method for achieving this level of accuracy.

sequence and 2.2 to 5.4 times smaller than the Halton sequence when $L \geq 2.5$. The QMC simulations using the Faure sequence tend to have the highest single sample error; nearly an order of magnitude greater than the Halton sequence. As mentioned in Section 6.2, the poor performance of the method may be attributed to the start-up error present in the Faure sequence, but more study is needed beyond this investigation for certainty.

In Figure 6.10, the critical error E_{crit} of the particle simulations is plotted for all the duct geometries tested in this section. As noted, the critical error E_{crit} is simply the error level at which the QMC simulation becomes a faster technique than the test particle Monte Carlo method. Since the QMC simulations have a higher error convergence rate and the cost of generating each sample trajectory varies little throughout the simulation, it is expected to reach all error levels smaller than E_{crit} in less time than the test particle Monte Carlo method. Hence, larger values of E_{crit} for a given QMC simulation indicate that it is the faster particle method for a wider

range of desired simulation accuracies. Overall, the value of E_{crit} generally decreases as L increases, which implies that the QMC simulations lose some of their effective range as the duct narrows. That is not to say that the QMC simulations are unable to outperform the test particle Monte Carlo method Monte Carlo method when L is larger. Rather, it means their performance gains are limited to simulations requiring a greater amount of accuracy.

Figure 6.10 demonstrates that the largest values of the critical error E_{crit} are found for the QMC particle simulations using the BCF-3 sequence ($L \leq 7.5$), and using the Niederreiter ($b = 2$) sequence ($L > 7.5$). For $L \leq 6$, the QMC simulations using the BCF-3 sequence outperform the test particle Monte Carlo method when the desired simulation accuracy is as high as 18%. In these cases, the QMC simulations using the BCF-3 sequence obtain large values of E_{crit} because their single sample error is small. In fact, the single sample error is so small for the QMC simulations using the BCF-3 sequence that they outperform the test particle Monte Carlo method after generating just one sample. For narrower duct geometries (*i.e.* $L > 7.5$), the QMC particle simulations using the Niederreiter ($b = 2$) sequence is the fastest particle method to achieve a simulation error less than 0.1%. In these cases, the superlative performance using the Niederreiter ($b = 2$) sequence is a result of the high error convergence rate of the method and the low computational cost needed to generate the sequence. Consequently, the critical error E_{crit} of the QMC simulations using the Halton sequence is smaller than the Niederreiter ($b = 2$) sequence but still remains within an order of magnitude of the Niederreiter ($b = 2$) sequence. As noted, the error of the QMC simulations using the Halton sequence converge at nearly the same rate as the Niederreiter ($b = 2$) sequence; however, they suffer from a longer computation time because of the greater costs of generating the Halton sequence.

In rather stark contrast, the QMC simulations using the Faure sequence have the smallest values of E_{crit} in Figure 6.10. Furthermore, the rate at which E_{crit} declines is much faster than the other QMC simulations because of the larger single sample error and greater computational cost associated with the Faure sequence. The important point to remember for all duct geometries tested in Figure 6.10 is that at least one of the QMC particle simulations using the Halton, BCF-3, and Niederreiter ($b = 2$) sequences is able to achieve a simulation accuracy of 0.1% in less time than the test particle Monte Carlo method.

While the critical error E_{crit} indicates the accuracy range where the QMC simulations are faster than the test particle Monte Carlo method, it does not give the magnitude of the actual performance gain of the QMC methods. In order to measure the performance gain, a reference error must be selected to compare the number of samples and computation time required by both the QMC simulations and the test particle Monte Carlo method. In this section, the reference error is chosen to be the expected relative error of the test particle Monte Carlo method after generating $N = 2^{23}$ samples. Specifically, this reference error varies between 10^{-4} and $5 \cdot 10^{-4}$, as shown in Figure 6.7. Using the power law models found for the error convergence, the number of samples and the computation time required by the QMC simulations to achieve the reference error are found and then normalized by the corresponding values for the test particle Monte Carlo method. From the normalized results, the factor by which the QMC simulations reduce the number of sample trajectories required to achieve the reference error is given in Figure 6.11(a), and the factor by which the QMC simulations speedup the computation time is given in Figure 6.11(b).

Consistent with the other performance metrics discussed in this section, the sample size reduction factor and the computation time speedup of the QMC particle

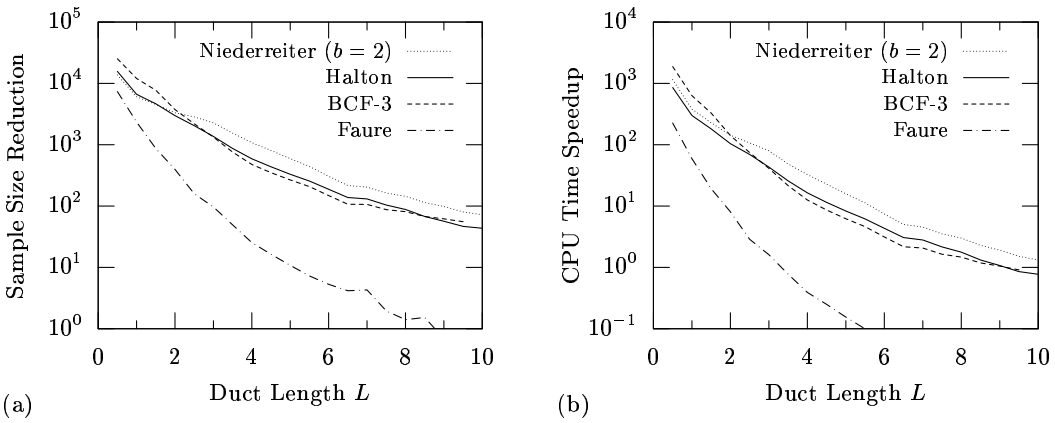

Figure 6.11: Performance gains of the QMC particle simulations when compared to the expected error of the traditional test particle Monte Carlo method after $N = 2^{23}$ samples: (a) the reduction factor in the number of sample trajectories needed by the QMC simulation; and (b) the speedup factor in the computation time of the QMC simulation.

simulations decline as the duct length to height ratio L increases. For the moment, exclude the results using the Faure sequence and only consider the performance gains for the other three QMC simulations using the Halton, BCF-3, and Niederreiter ($b = 2$) sequences. These QMC simulations are able to achieve the same reference error as the test particle Monte Carlo method by using over 10^4 times fewer sample trajectories when $L = 0.5$. While the the sample size reduction factor decreases as the duct narrows, these QMC simulations are still able to achieve the same reference error using 40 to 70 times fewer sample trajectories when $L = 10$. Since the computational cost of each sample trajectory is approximately constant during each QMC simulation, it is not surprising that the behavior of computation time speedup is similar to the sample size reduction factor in Figure 6.11. In particular, for the $L = 0.5$ duct geometry, the time required for these QMC simulations to reach the same reference error is nearly 10^3 times faster than the test particle Monte Carlo method. For the $L = 10$ duct geometry, the QMC particle simulation using the

Niederreiter sequence in base 2 is only 32% times faster than the test particle Monte Carlo method for reaching the reference error. As noted, the QMC particle simulations using the Halton and BCF-3 sequences are slightly slower than the Niederreiter ($b = 2$) sequence because of the higher construction costs. As a consequence, the QMC particle simulations using the Halton and BCF-3 sequences are also slightly slower than the test particle Monte Carlo method by a factor of 22% and 9% respectively when $L = 10$. However, the QMC particle simulations using the Halton and BCF-3 sequences are still able to achieve the reference error faster than the test particle Monte Carlo method for all duct geometries with $L \leq 9$. It is possible, by adopting the hybrid QMC/MC technique presented in Section 6.5, to further improve on the computation speedup attained by these QMC simulations (which can reduce the overall computation time by an additional factor of 2 to 4.5). Returning to the QMC simulation using Faure sequence, the performance gains are the lowest of all the QMC particle simulations. In fact, the QMC simulation using the Faure sequence requires fewer samples than the test particle Monte Carlo method when $L \leq 8.5$. Even worse, the QMC simulation using the Faure sequence is only able to reach the same reference error faster than the test particle Monte Carlo method when $L \leq 3$.

In summary, there are two main points demonstrated by the detailed geometry study presented in this section. First, for many of the duct geometries tested here, the QMC particle simulations using the Halton, BCF-3, and Niederreiter ($b = 2$) sequences achieve significant performance gains over the traditional test particle Monte Carlo method. In particular, the error convergence rate of all the QMC particle simulations tested remains superior to the $\mathcal{O}\left(N^{-1/2}\right)$ convergence rate of the Monte Carlo methods. More specifically, the QMC particle simulations using the Halton

and Niederreiter ($b = 2$) sequences achieve a near linear error convergence rate that is greater than $\mathcal{O}\left(N^{-0.85}\right)$ when $L \leq 3$. As a direct consequence of the higher convergence rate, the QMC simulations using the BCF-3 and Niederreiter ($b = 2$) sequences are faster than the test particle Monte Carlo method when the desired simulation error is less than or equal to 1% for the $L \leq 7.5$ duct geometries. For a given simulation error, the computation time of the QMC simulations can be orders of magnitude faster than the test particle Monte Carlo method. Furthermore, the computational speedup of the QMC simulations increases as the desired accuracy increases, which, as mentioned before, is a desirable feature when simulating low speed micro-scale flows.

The second main point of this section is that the performance gains of the QMC particle simulations tend to decline as the duct narrows; *i.e.* when the duct to length ratio L increases. The cause of this performance loss is attributed to the increase in the problem dimension that accompanies an increase in L, and the accompanying increase in the dimension of the low-discrepancy sequences needed for the QMC simulations as well. This problem of dimensionality of the low-discrepancy sequences is well-noted throughout literature for a wide range of QMC applications (see [23, 74, 114, 116, 117, 118, 120, 153, 167]). In particular, as the dimension of the low-discrepancy sequences increases, the presence of non-physical particle behavior in the QMC simulations developed here increases as well. The dimension problem is discussed in greater detail in Section 6.4. Despite the decline in performance, the QMC particle simulations using the Halton, BCF-3, and Niederreiter ($b = 2$) sequences still achieve non-trivial gains over the test particle Monte Carlo method when $L \leq 5$, which corresponds to a maximum problem dimension of $s = 124$. Even for the narrowest duct, when $L = 10$ and the problem dimension $s = 310$, the

QMC simulation using the Niederreiter sequence in base 2 is still faster than the test particle Monte Carlo method for reaching a simulation error less than 0.1%. Based on the overall performance demonstrated in the duct geometry study of this section, however, it seems reasonable to recommend a maximum low-discrepancy sequence dimension $s \approx 100$ as a design limit when developing new QMC particle simulations.

6.4 Correlation between dimensions of the low-discrepancy sequences

It is evident from the results in Sections 6.1, 6.2, and especially 6.3, that the performance of the QMC particle simulations suffers as the duct length to height ratio increases. As mentioned briefly in these previous sections, the decline in performance is attributed to an increase in the dimension of the low-discrepancy sequence needed by the QMC particle simulation. The dimensionality problem with the low-discrepancy sequences is well-noted throughout the literature for many different QMC applications, please see [23, 74, 114, 116, 117, 118, 120, 146, 153, 167]. In most cases, the dimensionality problem is only discussed in terms of the impact on the theoretical convergence rate of the discrepancy of the sequences; instead of the impact on the actual QMC simulation. Specifically, the growth and large magnitude of the implied constant in the asymptotic bound $\mathcal{O}\left(N^{-1}(\log N)^s\right)$ on the discrepancy are most often cited as a consequence of increasing the dimension of a low-discrepancy sequence. Also, Morokoff and Caflisch in [116] note that the minimum sequence length required before the asymptotic term $N^{-1}(\log N)^s$ begins to monotonically decrease also grows with the number of dimensions s of the low-discrepancy sequence.

These effects negatively impact the theoretical convergence rate for the star discrepancy of a low-discrepancy sequence; and, by application of the Koksma-Hlawka

inequality in (3.5), they also increase the upper error bound on the QMC method. However, as mentioned throughout this investigation, the Koksma-Hlawka inequality does not provide a very tight bound on the error of the QMC method for most applications. An unfortunate consequence of this fact is that there does not seem to be any clear connection between the dimension problems associated with the theoretical error convergence, and the actual loss of physical accuracy of the QMC simulation observed in practice. Thus, it is the goal of this section to discuss the loss of the physical accuracy in terms of the non-physical correlation that persists between the dimensions of the low-discrepancy sequences used for the QMC particle simulations. This is not an entirely new concept; Morokoff and Caflisch [116], and Press and Teukolsky [146] graphically illustrate some of the correlation patterns present between the pairs of dimensions taken from the Sobol', Halton, and Faure sequences. However, this section takes the concept much farther by providing an actual estimate of the magnitude and extent of the correlation that exists between the dimensions of the four main low-discrepancy sequences used in this investigation.

For the moment, consider what happens in the actual QMC particle simulation when two consecutive dimensions of the low-discrepancy sequence, x_1 and x_2, are highly correlated; that is, $x_1 \approx x_2$. In such a case, the two particle moves, or wall collisions, generated by x_1 and x_2 will behave very similarly. Both particle moves will almost always be in the same direction with nearly the same trajectory, which is essentially the same behavior that occurs when the second wall collision is a specular reflection. However, as stated in the initial assumptions given in Section 5.1 for the free molecular flow simulation, the interior duct walls are fully diffuse. Thus, the nearly specular wall collisions, which occur in the QMC simulation when two consecutive dimensions of the low-discrepancy sequence are highly correlated, are

not physically consistent with the actual problem being simulated.

Furthermore, this loss of physical accuracy of the QMC simulation is not limited to just the correlation between consecutive dimensions of the low-discrepancy sequence, correlation between any two dimensions is physically inconsistent as well. Recall from the development of the QMC particle simulation in Section 5.5 that each dimension of the low-discrepancy sequence is used to generate the trajectory path after a particle undergoes a collision with the diffuse duct wall. One of the characteristics of a collision with a diffuse wall is that the trajectory of the particle leaving the wall is completely independent of any prior wall collisions. Basically, a simulated particle undergoing a diffuse wall collision must behave as if it lost all "memory" of any previous collisions along its trajectory path in order to be physically consistent. Now if any two dimensions of the low-discrepancy sequence, x_i and x_j (with $i < j$), are highly correlated such that $x_i \approx x_j$, then the j^{th} collision depends on the i^{th} collision; which implies that the simulated particle retains some memory of its earlier trajectory. Therefore, correlation between any two dimensions of the low-discrepancy sequence used in the QMC method is not physically consistent with the problem being simulated.

Before proceeding with the discussion on correlation, some clarification is needed to distinguish between the two types of correlation mentioned in this investigation for low-discrepancy sequences. The first type of correlation is the kind that naturally occurs between the elements of a one dimensional low-discrepancy sequence. If one fails to recognize that the elements of a one dimensional sequence are highly correlated, and treats each element as an independent sample from a uniform distribution, the results can be grossly inaccurate. An example of this mistake is presented in Section 5.5, where a physically inconsistent particle simulation is obtained when the pseudo-

random number generator is replaced with the van der Corput sequence. While this type of correlation may cause problems when not taken into account, there are actually some advantages to the correlation. As noted by Press and Teukolsky [146], each new point added to the low-discrepancy sequence effectively "knows" where all the previous sequence points are located; and as such, the new point is placed in a position that "maximally avoids" these other points. It is by virtue of this natural correlation that the low-discrepancy sequence is able to efficiently achieve an even distribution of points throughout the unit interval.

In order to avoid this first type of correlation, multi-dimensional low-discrepancy sequences are used, when independent samples from a uniform distribution are needed. Each dimension of a multi-dimensional low-discrepancy sequences is actually a one dimensional low-discrepancy sequence generated from a unique constructive element.[14] In order for the dimensions of a multi-dimensional low-discrepancy sequence to be independent, these constructive elements must also be independent in some sense. More specifically, the constructive elements are independent under the following conditions: when the bases of the Halton sequence are pair-wise relatively prime, when the irrational numbers of the BCF-3 sequence are linearly independent over $\mathbb{Q}$, and when the polynomials used to generate the Niederreiter sequence in base 2 are irreducible over $\mathbb{F}_2[x]$. If the constructive elements of a multi-dimensional low-discrepancy sequence are independent, then it is physically accurate to treat each dimension as an independent sample from a uniform distribution. This physical accuracy is made certain by the convergence of the Koksma-Hlawka inequality; unfortunately, it is only guaranteed in the limit as the sequence length tends to-

[14]Here "constructive element" refers to the mathematical object that governs the construction of a low-discrepancy sequence. For the Halton sequence, it is a positive integer. For the Weyl-Richtmyer sequence, it is an irrational number. For the Faure sequence, it is a degree one polynomial over a prime field. And for the Niederreiter sequence in base 2, it is a polynomial in $\mathbb{F}_2[x]$.

ward infinity. Thus, it is entirely possible for two independent dimensions of a low-discrepancy sequence to behave as if dependent over a finite sequence length.

The second type of correlation, and the focus of this section, is the kind that occurs when two dimensions of a low-discrepancy sequence appear dependent throughout the sequence lengths used in practice for the QMC particle simulation. When this occurs, the resulting QMC particle simulation is not producing a physically consistent representation of the diffuse wall collision process. As with the first type of correlation, the second type can not be eliminated from a multi-dimensional low-discrepancy sequence because it is what actually enables the sequence to efficiently achieve a even distribution of points throughout the domain. The sequence lengths over which the dimensions of the low-discrepancy sequence appear dependent on each other is then the critical feature of this type of correlation. The Koksma-Hlawka inequality guarantees that any correlation between the dimensions of a low-discrepancy sequence is eventually broken up. The question then becomes how long does it take before the sequence appears uncorrelated. Intuitively, if the correlation only persists over sequence lengths that are much smaller than those used by the QMC simulation, then the effect of the second type of correlation is expected to be negligible. Unfortunately, as the dimension of the low-discrepancy sequence increases, so too does the sequence length over which the dimensions appear correlated. Even though each dimension of a low-discrepancy sequence is generated from a unique constructive element, these constructive elements become more similar as the number of dimensions increases. As the differences between the constructive elements become smaller, their behavior generating the dimensions of the low-discrepancy sequence becomes more similar. A greater sequence length is therefore required before the small differences between the dimensions grow large enough to break up the apparent correlation.

The purpose of this section is to identify which dimensions of the four main low-discrepancy sequences used for the QMC particle simulation are likely to have significant correlation, and to illustrate the construction patterns caused by the correlated dimensions. In order to evaluate the impact of the correlation caused by the construction pattern, a measure of the correlation magnitude is introduced here. In addition, the extent, or persistence, of the correlation caused by these construction patterns is measured by the sequence length necessary to reach a correlation level equivalent to a random sequence of the same length. The first half of this section uses the Halton sequence as an example to develop and calculate these measures for the magnitude and extent of the correlation present in the low-discrepancy sequences. The second half of this section calculates and compares the magnitude and extent of the correlation found in the BCF-3, Faure, and Niederreiter ($b = 2$) sequences.

6.4.1 Correlation between two dimensions of the Halton sequence

To begin the discussion of correlation between dimensions of a low-discrepancy sequence, consider the case of a two dimensional Halton sequence $S_H = \mathbf{x}_0, \mathbf{x}_1, \ldots$ where $\mathbf{x}_n = \big(\chi_{p_1}(n), \chi_{p_2}(n)\big)$ for all n. Specifically, let the prime bases p_1 and p_2 be large twin primes such that $p_2 = p_1 + 2$. With these prime bases, the first p_2 elements of the Halton sequence are given by

$$(\mathbf{x}_0, \ldots, \mathbf{x}_{p_2-1}) = \big((0,0), (\tfrac{1}{p_1}, \tfrac{1}{p_2}), (\tfrac{2}{p_1}, \tfrac{2}{p_2}), \ldots, (\tfrac{1}{p_1^2}, \tfrac{p_2-2}{p_2}), (\tfrac{p_1+1}{p_1^2}, \tfrac{p_2-1}{p_2})\big). \tag{6.2}$$

Note that the first p_1 elements of this Halton sequence fall nearly on the line $y = x$ in the Cartesian plane, which indicates a high degree of correlation between the dimensions of the subsequence in (6.2). In fact, if one recalls the algorithmic implementation of the Halton sequence from Section 4.3, most of the sequence elements can be determined explicitly by adding a constant to each dimension of the previous

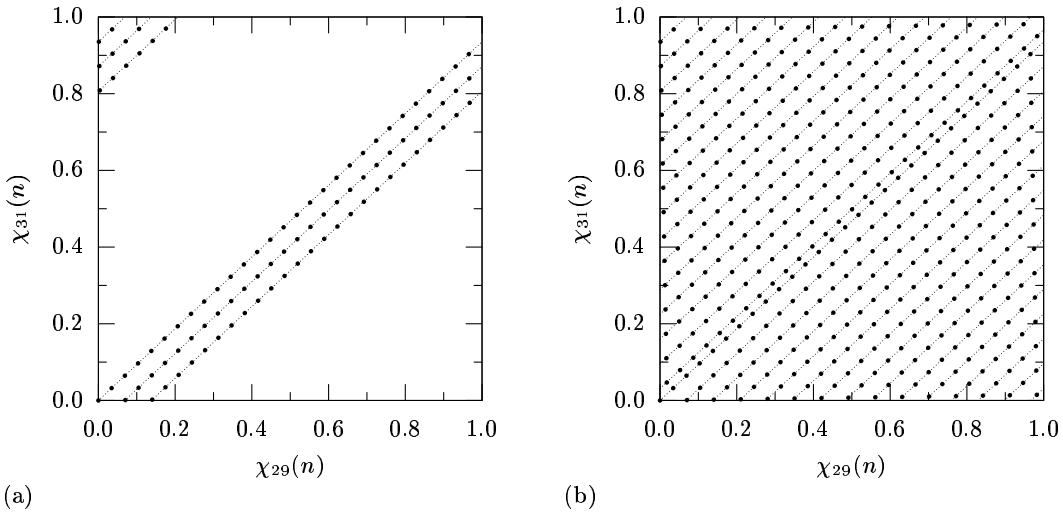

Figure 6.12: Plot of the two dimensional Halton sequence in prime bases $p_1 = 29$ and $p_2 = 31$: (a) for $0 \leq n < 3p_2$; and (b) for $0 \leq n < p_2\lfloor\frac{p_1}{p_2-p_1} + 1\rfloor$.

element. That is, $\mathbf{x}_n = \mathbf{x}_{n-1} + (\frac{1}{p_1}, \frac{1}{p_2})$ for all $n > 0$ except when $n \equiv 0$ modulo p_1 or p_2. Thus, consecutive elements of this Halton sequence appear on the same line in the Cartesian plane with a slope of $\frac{p_1}{p_2}$. An example of this distribution pattern is given in Figure 6.12 for a Halton sequence in prime bases $p_1 = 29$ and $p_2 = 31$.

Now consider consecutive blocks of p_2 elements from this Halton sequence. In the first block, $n = 0, \ldots, p_2 - 1$, the first p_1 points of the Halton sequence appear on the line through the origin $y = \frac{p_1}{p_2}x$, with the last two points of (6.2) appearing near the $(0, 1)$ corner of the unit square, as shown in Figure 6.12(a). In the second block, $n = p_2, \ldots, 2p_2 - 1$, the first $p_1 - 2$ elements of this block appear on a parallel line $y = \frac{p_1}{p_2}(x - \eta)$ shifted to the left of the origin along the $x-$axis by the amount

$$\eta = \frac{2}{p_1}\left(1 + \frac{1}{p_1 p_2}\right). \tag{6.3}$$

Similar to the first block, the last four elements of the second block wrap around the unit square and appear near the $(0, 1)$ corner along a parallel line shifted by the same amount η. The pattern continues for each successive block of p_2 elements from the

Halton sequence, shifting the parallel lines by the same distance along the $x-$ axis. Each new line added for $y < \frac{p_1}{p_2}x$ has two fewer points than the previous line, while the continuation of the same line for $y > \frac{p_1}{p_2}x$ has two more points than the previous line. In general, the change in the number of points along each successive line is equal to $p_2 - p_1$ when $p_2 - p_1 \ll p_1, p_2$. At the start of the $(\lfloor \eta^{-1} \rfloor + 1)^{th}$ block of p_2 elements, the parallel line is so close to the corner $(1, 0)$ that there is only room for a single point of the sequence. The remaining elements of the block wrap around the unit square to continue along a parallel line that is close to the first block of elements. In fact, at this point in the construction of the Halton sequence, the points in the $(\lfloor \eta^{-1} \rfloor + 1)^{th}$ block have the closest spacing to any of the previous sequence elements, as illustrated in Figure 6.12(b).

As a consequence of constructing points in this manner, the Halton sequence is evenly distributed for the first $p_2 \lfloor \eta^{-1} \rfloor$ elements. Given two sets of points, the more evenly distributed point set yields a smaller overall correlation between the dimensions. Hence, the first $p_2 \lfloor \eta^{-1} \rfloor$ elements are expected to possess, when viewed as complete set, a very small amount of correlation between the dimensions. However, for subsequences that are smaller than $p_2 \lfloor \eta^{-1} \rfloor$ elements, the points generated by this Halton sequence are highly correlated. In more general terms, there are two scales of behavior for this type of Halton sequence. If one inspects a relatively small number of consecutive elements of a low-discrepancy sequence, one should expect to find a high degree of correlation between the dimensions. In contrast, for a sufficiently large number of consecutive elements of a low-discrepancy sequence, one should find a negligible degree of correlation. It is important to note that these two scales of correlation behavior are present in the construction of all low-discrepancy sequences.

For the example of the two dimensional Halton sequence using the large twin

prime bases, the subsequence length of $p_2\lfloor\eta^{-1}\rfloor$ serves as a natural transition point between these two scales. Each subsequent block of elements from the Halton sequence, with length $p_2\lfloor\eta^{-1}\rfloor$, covers the unit square with the same pattern shown in Figure 6.12(b) for the first $p_2\lfloor\eta^{-1}\rfloor$ elements. Similar to the smaller blocks of p_2 elements considered earlier, the larger blocks of length $p_2\lfloor\eta^{-1}\rfloor$ are shifted from the first block by an even amount smaller than η in (6.3). Since each block of $p_2\lfloor\eta^{-1}\rfloor$ elements is uniformly distributed throughout the unit square, the Halton sequence achieves a minimal amount of correlation when the sequence length is a multiple of $p_2\lfloor\eta^{-1}\rfloor$. While the construction of the Halton sequence appears to yield a cyclic correlation between the dimensions with a period of $p_2\lfloor\eta^{-1}\rfloor$, it is important to note that all low-discrepancy sequences are actually infinite, non-repeating sequences. Hence, the correlation behavior and construction patterns are not truly cyclic; and as such, will be referred to as near-cyclic in this investigation. It is important to note that the near-cyclic behavior of this Halton sequence is not limited to just the case of large twin primes. Similar correlation behavior and construction patterns are found in a more general case; that is, when two prime bases of the Halton sequence satisfy $|p_1 - p_2| \ll p_1, p_2$. Following the same reasoning used to derive (6.3), and ignoring terms that are relatively small; the expected period of the near-cyclic correlation of this more general Halton sequence is approximately $\frac{p_1 p_2}{|p_1 - p_2|}$.

The period of the near-cyclic correlation present in a low-discrepancy sequence is useful for assessing the physical accuracy of the QMC simulation that uses the sequence. In an ideal setting, one could achieve great simulation accuracy by simply stopping the simulation at a low-discrepancy sequence length that corresponds to a point of minimum magnitude in the correlation near-cycle. However, it is not possible to accomplish this in practice because each pair of dimensions in a multi-dimensional

low-discrepancy sequence has a different period for the near-cyclic correlation. These differences can span many orders of magnitude, which makes finding an ideal sequence length at the minimum point of every correlation near-cycle impractical. Instead, if one wants to ensure that the correlation between dimensions is physically negligible, one must select a sequence length of the low-discrepancy sequence that is sufficiently longer than the maximum period of the correlation near-cycle for all dimension pairs. As the length of the low-discrepancy sequence extends beyond the period of the maximum near-cycle, the effect of the local subsequence correlation on the overall correlation becomes smaller. In order to assess if the correlation between dimensions of a low-discrepancy sequence is sufficiently small, it is useful to formally introduce the statistical correlation in two dimensions as a measure. For a sequence with length N, the statistical correlation between two dimensions (x_1, x_2) of the sequence is denoted by $\tilde{\rho}_{12}$ which is defined by

$$\tilde{\rho}_{12}(N) = \frac{\overline{x_1' x_2'}}{\left(\overline{x_1'^2} \cdot \overline{x_2'^2}\right)^{1/2}} \tag{6.4}$$

where the over bar denotes an average quantity, and the prime denotes the difference in the quantity from the sample mean.

Rather than use the sample mean and sample variance in the statistical correlation in (6.4), it is more convenient to use the expected mean of the sequence $\overline{x_1} = \overline{x_2} = \frac{1}{2}$, and the expected variance $\overline{x_1'^2} = \overline{x_2'^2} = \frac{1}{12}$. The modification serves to simplify the resulting analysis without changing the utility of the correlation measure. This modified statistical correlation is denoted by ρ_{12}, and is defined by

$$\rho_{12}(N) = 12 \sum_{n=1}^{N} \left(x_{1,n} - \tfrac{1}{2}\right)\left(x_{2,n} - \tfrac{1}{2}\right). \tag{6.5}$$

For the remainder of the investigation, the modified statistical correlation ρ_{12} in (6.5) is simply referred to as the two dimensional correlation. When $\rho_{12} = 0$, the dimen-

sions of the low-discrepancy sequence are said to be uncorrelated, which indicates the QMC method using the sequence yields a good approximation of the physical process being simulated. Alternatively, when $\rho_{12} > 0$, the dimensions of the low-discrepancy sequence are said to be positively correlated; and when $\rho_{12} < 0$, the dimensions of the low-discrepancy sequence are said to be negatively correlated. The traditional definition of the correlation $\tilde{\rho}_{12}$ in (6.4), based on the sample means and variance, is strictly within the interval $[-1, 1]$. Here $\tilde{\rho}_{12} = \pm 1$ indicates perfect correlation (positive or negative) between the two dimensions. However, for the simplified definition of ρ_{12} in (6.5), it is possible that the two dimensional correlation may be slightly outside the interval $[-1, 1]$.

After establishing ρ_{12} in (6.5) as a measure of the correlation present between dimensions of a low-discrepancy sequence, it is natural to ask, "what values of ρ_{12} would indicate an uncorrelated sequence?" While there is no definite answer, it is useful to compare the correlation of a low-discrepancy sequence to that of a random sequence. The dimensions of a random sequence, by definition, should be uncorrelated; that is to say, the expected value of ρ_{12} is zero. However, the expected value of the correlation is almost never the actual value measured given a finite sequence generated at random. From the central limit theorem, the correlation ρ_{12} of a finite sequence generated at random is bound with a fixed probability; and the bound decreases as $\mathcal{O}\left(N^{-1/2}\right)$ with the sequence length N. Specifically for the 95% confidence interval, the bound on the correlation of a random sequence of N elements in $[0, 1)$ is given by

$$\text{Prob}\left[|\rho_{12}(N)| \leq \frac{1.39}{\sqrt{N}}\right] \approx 0.95. \tag{6.6}$$

For convenience, let ρ_{mc} denote this bound on the correlation of the random sequence, *i.e.* $\rho_{mc} = 1.39/\sqrt{N}$.

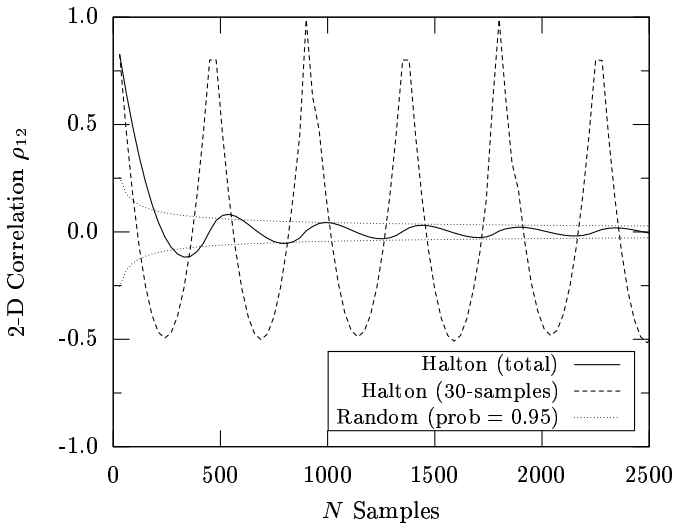

Figure 6.13: Comparison of the running and local two dimensional correlation ρ_{12} for the Halton sequence $\mathbf{x}_n = (\chi_{29}(n), \chi_{31}(n))$.

In order to illustrate the near-cyclic behavior of the correlation between dimensions of the Halton sequence, the local two dimensional correlation is found for the previous example $\mathbf{x}_n = \big(\chi_{29}(n), \chi_{31}(n)\big)$, see Figure 6.12. In particular, the two dimensional correlation ρ_{12} is plotted in Figure 6.13 for consecutive blocks of 30 elements of the Halton sequence. The correlation between these 30 elements blocks is often much stronger than expected for a random sequence of 30 elements, that is, $\rho_{mc} \approx 0.25$. The behavior of the local correlation clearly appears cyclic, with a period that corresponds very well to the estimate, $\frac{p_1 p_2}{|p_1 - p_2|} \approx 450$, provided earlier in this section. Consequently, the running correlation of ρ_{12} for the overall sequence length is approximately zero when the sequence length is a multiple of the period of local correlation. Moreover, as the sequence length increases, the impact of the local correlation on the overall correlation is shown to steadily decline as expected. In fact, after the completion of the second near-cycle, it appears that the correlation ρ_{ld} between the dimensions of the Halton sequence remains within the 95% confidence

interval for the correlation ρ_{mc} of a random sequence. Thus, for sequence lengths greater than $N \approx 1000$, one can state that the two dimensions of this specific Halton sequence are at least as uncorrelated as a random sequence.

If one is able to state that the dimensions of a specific low-discrepancy sequence are at least as uncorrelated as a random sequence of equal length, then there is much greater confidence that the QMC method using the low-discrepancy sequence is accurately simulating the physical behavior. The calculation of the running correlation ρ_{ld} for a low-discrepancy sequence is equivalent to the QMC integral approximation of the function $f(x_1, x_2) = 12(x_1 - \frac{1}{2})(x_2 - \frac{1}{2})$. Given that

$$\int_{\bar{I}^2} f(x_1, x_2) dx_1 dx_2 = 0,$$

and the integrand is of bounded variation in the sense of Hardy and Krause, with $V_{HK}(f) = 24$; the Koksma-Hlawka inequality (3.5) provides a bound on the running correlation ρ_{ld}. Specifically,

$$|\rho_{ld}(N)| \leq 24 D_N^*(\mathbf{x}_1, \ldots, \mathbf{x}_N),$$

where the star discrepancy of the low-discrepancy sequence converges to zero as $\mathcal{O}\left(N^{-1}(\log N)^2\right)$. In contrast, the upper bound on the confidence interval for the random sequence correlation ρ_{mc} converges more slowly as $\mathcal{O}\left(N^{-1/2}\right)$. As a result of the faster convergence of the low-discrepancy sequence, there must exist a sequence length N_{min} for which the running correlation of the low-discrepancy sequence remains within the 95% confidence interval (6.6) for the correlation of a random sequence. Stated more formally,

$$N_{min} = \min\{N \ : \ \rho_{ld}(n) < \rho_{mc}(n) \text{ for all } n > N\}. \tag{6.7}$$

Thus, for low-discrepancy sequences with a length greater than N_{min}, one is able to state that the dimensions of a specific low-discrepancy sequence are at least as

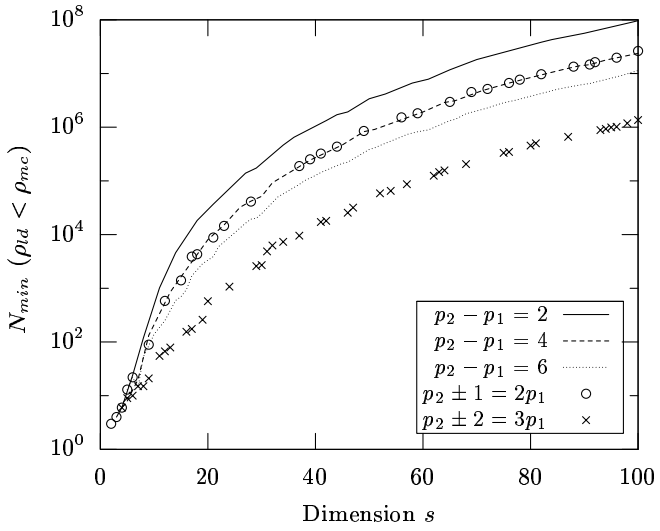

Figure 6.14: The minimum sequence length N_{min} required for the Halton sequence constructed with prime bases p_1 and p_2 to be considered as uncorrelated as a random sequence.

uncorrelated as a random sequence of equal length. It is interesting to point out that this correlation check of the low-discrepancy sequence is very similar to the serial test for pseudo-random number generators [78].

In practice, the Halton sequence in s dimensions is constructed using the smallest distinct primes $p_1, \ldots, p_s$ as the bases of the van der Corput sequences used to construct each dimension. Naturally, it follows that the maximum period of the correlation near-cycle due to the twin prime bases also increases with dimension of the Halton sequence.[15] In Figure 6.14, N_{min} is found for the largest twin prime bases that are used to construct a Halton sequence in s dimensions. It is not surprising then to see an increase in the minimum sequence length N_{min} needed for the dimensions generated by the twin prime bases to appear as uncorrelated as a random

[15]For the dimension sizes of the low-discrepancy sequences needed throughout this investigation, a significant fraction of the prime bases happen to belong to a twin prime pair. Specifically, there are 16 twin prime pairs among the first 50 primes (64% of all bases); there are 25 twin prime pairs among the first 100 primes (50% of all bases); and, there are 60 twin prime pairs among the first 300 primes (40% of all bases). In fact, as asserted by the Twin Prime Conjecture, there is most likely an infinitude of twin prime pairs [185].

sequence. It is interesting to note that the increase in N_{min} is not solely the result of an increase in the period of the correlation near-cycle. There is also an increase in the number of near-cycles needed for the correlation of the low-discrepancy sequence to reach the level of a random sequence. For the previous example of the Halton sequence using prime bases $p_1 = 29$ and $p_2 = 31$, given in Figure 6.12, less than 2.3 near-cycles ($N \approx 10^3$) are needed before the correlation between these two dimensions remains with the bound $\rho_{mc}(N)$ for the random sequence. In contrast, for the Halton sequence using prime bases $p_1 = 521$ and $p_2 = 523$, which is the largest twin prime pair used to construct a 100 dimensional sequence, over 650 near-cycles ($N \approx 10^8$) are necessary to reduce the correlation to the level of $\rho_{mc}(N)$.

It is rather unsettling to observe in Figure 6.14, that for a Halton sequence with more than 64 dimensions, the correlation between some dimension pairs persists for sequence lengths greater than the $N = 2^{23}$ samples used for the length study in Section 6.3. However, as noted by Morokoff and Caflisch in [116], the mere presence of correlation in the low-discrepancy sequence does not necessarily condemn the QMC method to poor performance; much depends on the actual function being integrated. If a function is extremely sensitive to correlation (*i.e.* there exists a large inter-dependence between the dimensions), then sampling the function with a highly correlated low-discrepancy sequence is likely to yield poor convergence of the QMC approximation. For the QMC particle simulation using the Halton sequence, the largest prime bases are used to generate the last moves of the sample trajectory, when the simulated particle weight is the smallest. Consequently, the impact of the correlation between the higher dimensions on the sample trajectory's contribution to the conductance probability is much less.

For a free molecular duct geometry $L \leq 4$, which corresponds to a low-discrepancy

sequence dimension $s < 100$, the most persistent correlation in the Halton sequence occurs for trajectory moves when the particle weight is less than 10^{-6}. Thus, it is fair to assume that the effect of the two dimensional Halton sequence is negligible for the QMC particle simulation when $L \leq 4$. However, for the narrowest duct geometry in this investigation (*i.e.* $L = 10$), the most persistent two dimensional correlation from the first 100 dimensions occurs for trajectory moves where the particle weight is typically greater than 10^{-2}. Also, consider when the trend in Figure 6.14 is extrapolated to the 300 dimensional Halton sequences required for the QMC particle simulation of the $L = 10$ duct geometry. In this case, the maximum length of the correlation near-cycle between the largest twin prime bases is nearly 2 million sequence elements. Assuming that the number of near-cycles needed to reach N_{min} continues to increase as well, this correlation will persist over sequence lengths that are orders of magnitude longer than can be simulated in practice. As the duct geometry narrows (*i.e.* L increases), low-discrepancy sequences with larger dimensions are needed for the QMC particle simulations. Unfortunately, the magnitude and extent of non-physical correlation between the dimensions also increases. Therefore, as a consequence of the increase in the correlation between the dimensions of the low-discrepancy sequence, the QMC particle simulation of the free molecular duct yields a less accurate representation of the true diffuse wall collision processes that govern the flow.

Since correlation between the dimensions of the low-discrepancy sequence can negatively impact the physical accuracy of the QMC particle simulation, it is natural to consider if it is possible to reduce the correlation by choosing an alternate sequence construction. An obvious first attempt for the Halton sequence would be to simply remove one-half of each pair of bases that are twin primes; thus, elim-

inating the correlation problem caused by these dimensions. Unfortunately, there are two problems to such a strategy. First, removing any of the s smallest primes $p_1, \ldots, p_s$ requires that the replacement prime bases are larger than the originals. As a consequence, the bound on the star-discrepancy increases for the Halton sequence. Moreover, the one dimensional projection of the sequence in the dimensions generated from the larger replacement prime bases is not as evenly distributed as before. Second, even if there are no twin prime pairs used to generate the dimensions of the Halton sequence, there are many other combinations for the prime bases that yield significant correlation between the dimensions.

Recall that the construction pattern in Figure 6.12 is not limited to just twin primes; in fact, it occurs whenever $p_1 \approx p_2$ and $|p_1 - p_2| \ll p_1, p_2$. Thus, the cousin primes (*i.e.* those of the form $p_2 - p_1 = 4$), and the sexy primes[16] (*i.e.* those of the form $p_2 - p_1 = 6$), are also able to produce persistent correlation when used as the prime bases of the Halton sequence. The minimum sequence length N_{min}, used to measure the extent of the correlation between the dimensions of the low-discrepancy sequence, appears to increase at the same rate for the twin, cousin, and sexy prime pairs shown in Figure 6.14. The period of the near-cyclic correlation produced from these construction patterns is approximately equal to $\frac{p_1 p_2}{|p_1 - p_2|}$. Hence, the length of the correlation near-cycle for the cousin and sexy prime pairs is 2 and 3 times shorter, respectively, than for the twin prime pairs of the same magnitude. Consequently, for the first 100 dimensions of the Halton sequence, the value of N_{min} for the twin prime pairs is no more than 4 times larger than the value found for the cousin prime pairs of similar magnitude. Similarly, the value of N_{min} for the twin prime pairs is

[16]The terminology "cousin primes" and "sexy primes" appears in general discussions of prime numbers found in literature, *e.g.* [185], and are not due to the author. In particular, the salacious nomenclature for prime pairs of the form $p_2 - p_1 = 6$ is attributed to the Latin origins of the number six.

no more than 10 times larger than the value found for the sexy prime pairs of similar magnitude. While the extent of the two dimensional correlation is not as great as for the twin prime pairs, it is still quite significant between the cousin and sexy prime pairs. More specifically, for the largest cousin and sexy prime pairs used to construct the 100 dimensional Halton sequence, the correlation between these dimensions does not approach the same level as a random sequence until at least $N = 10^7$ elements have been generated.

Up until now, the only construction pattern considered for the Halton sequence is the case when the two prime bases are nearly the same magnitude, *i.e.* $p_1 \approx p_2$. For this specific case, the sequence points are generated along lines that are parallel with the line $x_2 = \frac{p_1}{p_2} x_1$, as illustrated earlier in Figure 6.12. However, this is not the only construction pattern in the Halton sequence that yields significant two dimensional correlation. In fact, significant two dimensional correlation can occur whenever prime bases of the form $p_2 = \frac{a}{b} p_1 \pm c$ are used in the Halton sequence; where p_1 and p_2 are relatively large prime numbers, and a, b, and c are relatively small integers with $\gcd(a, b) = 1$. One example of this type of construction pattern occurs when the dimensions of the Halton sequence are generated by a pair of Sophie Germain primes; that is, two primes p_1 and p_2 that share the property $p_2 = 2p_1 + 1$. When a pair of Sophie Germain primes are used as the bases of a two dimensional Halton sequence, the elements are generated along lines that are nearly parallel to the line $x_2 = \frac{1}{2} x_1 \pmod 1$. More generally, if two prime bases of the Halton sequence have the form $p_2 = \frac{a}{b} p_1 \pm c$, then the elements are generated along lines that are nearly parallel to the line $x_2 = \frac{b}{a} x_1 \pmod 1$. It is interesting to note, that the construction pattern for the Sophie Germain primes yields a correlation near-cycle which is approximately twice that of the cousin primes, when the largest prime of

each pair is the same magnitude. However, the magnitude of the local correlation is only half as large for the Sophie Germain primes. Consequently, for dimensions of the Halton sequence constructed from either a pair of cousin primes, or a pair of Sophie Germain primes, the extent of the correlation N_{min} between these dimension pairs is the nearly same, as indicated in Figure 6.14.

It is possible to select the prime bases used in the construction of the Halton sequence such that no two bases satisfy any of the conditions plotted in Figure 6.14. However, eliminating these prime pairs, which are all shown to produce significant correlation between the dimensions of the Halton sequence, forces the use of exceedingly large prime bases instead. While the extent of two dimensional correlation is reduced in this case, it comes at the high price of creating a poor distribution of elements in each one dimensional projection of the Halton sequence. Thus, for the Halton sequence, it does not seem practical to avoid the two dimensional correlation problems by simply removing some of the prime bases. In terms of the accuracy of the QMC particle simulation, the surest way to control the negative effects of a correlated low-discrepancy sequence is to collect enough samples such that the overall correlation of the sequence is on the same order as a random sequence. If a sufficient number of samples is collected, *i.e.* $N > N_{min}$, then the QMC particle simulation is expected to be a physically accurate approximation to the problem, at least with regards to capturing the correct behavior of the diffuse wall collisions.

6.4.2 Correlation between two dimensions of the BCF-3 sequence

It is important to note that the correlation problem is not just limited to the Halton sequence. In fact, as will be shown next, it is possible to find similarly persistent two dimensional correlation from the construction patterns of the other

three low-discrepancy sequences used in this investigation. For example, the BCF-3 sequence in Figure 6.15(a) shows a construction pattern that exists when the fractional parts of the irrational numbers used to generate the dimensions are very close together. This construction pattern is similar to that of the Halton sequence with twin prime bases; and as such, it is capable of producing significant correlation in the BCF-3 sequence as well. Before proceeding, it should be noted that the following analysis of the correlation present in the BCF-3 sequence applies to all Weyl-Richtmyer low-discrepancy sequences.

To discuss the correlation in the BCF-3 sequence more formally, consider two dimensions (x_1, x_2) of the sequence generated by the irrational numbers z_1 and z_2. Next, let δ denote the difference between the fractional part of the irrational numbers; hence,

$$\delta = \big|[z_1] - [z_2]\big|, \tag{6.8}$$

where the square brackets denote the fractional part of the argument, *i.e.* $[x] = x - \lfloor x \rfloor$. The first N elements of the BCF-3 sequence are then found to be evenly distributed in a narrow band along the line $x_1 = x_2$ when $N \ll \delta^{-1}$, as illustrated in Figure 6.15(a). The width of this band grows with the length N of the BCF-3 sequence; and if measured in either the x_1 or x_2 direction, the width of the band is equal to $\delta \cdot N$. In addition, the band of sequence points appears below the line $x_1 = x_2$ when $[z_1] > [z_2]$; and conversely, the band appears above the line when $[z_1] < [z_2]$.

As the length N of the BCF-3 sequence increases, the width of the band continues to increase until $N = \lfloor \delta^{-1} \rfloor$; at which point the band covers the entire domain. Furthermore, as the band widens, the points of the BCF-3 sequence still remain evenly distributed within the band. Thus, when $N = \lfloor \delta^{-1} \rfloor$, the BCF-3 sequence

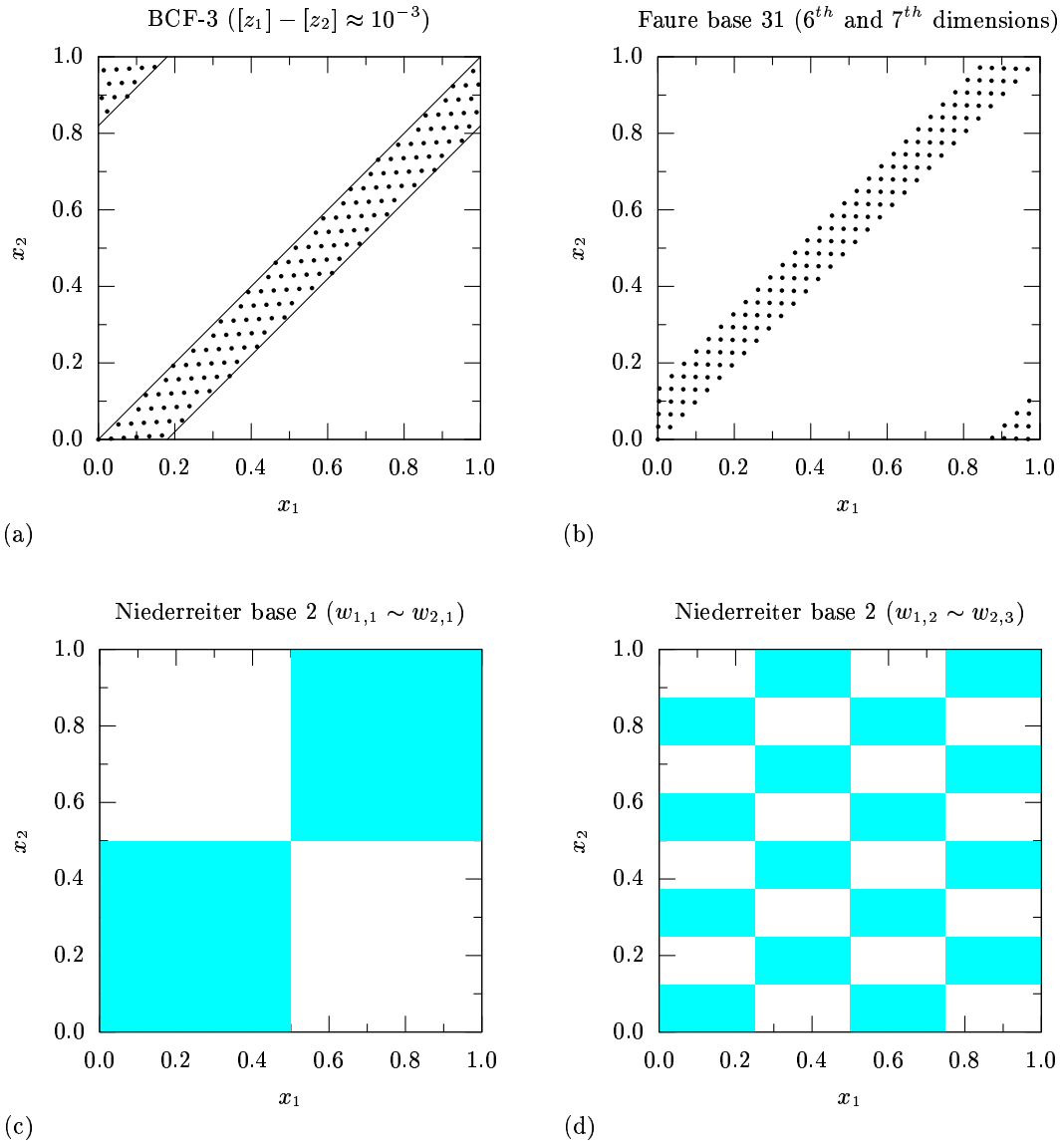

Figure 6.15: Specific examples of the two dimensional construction patterns that can produce significant correlation between the dimensions of the low-discrepancy sequences: (a) the BCF-3 sequence; (b) the Faure sequence in base 31; (c) the Niederreiter sequence in base 2 with the most significant bit of each dimension correlated; and (d) the Niederreiter sequence in base 2 with the 2^{nd} most significant bit correlated with the 3^{rd} most significant bit.

is evenly distributed throughout the unit square $[0,1)^2$, and the running correlation $\rho_{12}(N)$ between the two dimensions, as defined in (6.5), is nearly zero. The general construction pattern described in the first half of this section for the Halton sequence (see Figure 6.12) is essentially the same as the BCF-3 sequence when $\delta << 1$. Specifically, each successive block of $\lfloor \delta^{-1} \rfloor$ elements of the BCF-3 sequence repeats nearly the same construction pattern as the first block. The only difference is that each new block of $\lfloor \delta^{-1} \rfloor$ elements is translated by a small amount in order to prevent any duplicate sample points, and to continue the even distribution of the sequence elements throughout the domain. Therefore, based on the same reasoning previously used for the Halton sequence, the period of the correlation near-cycle of the two dimensional BCF-3 sequence is approximately $\lfloor \delta^{-1} \rfloor$, when $\delta \ll 1$.

For the construction pattern illustrated in Figure 6.15(a), the maximum length of the correlation near-cycle is the inverse of the closest distance between any two fractional parts used to construct a multi-dimensional BCF-3 sequence. Recall from Chapter IV that each dimension of the BCF-3 sequence is generated from the fractional part of a distinct irrational number, which is restricted to the unit interval $[0,1)$. Thus, as the dimension of the BCF-3 sequence increases, these distinct fractional parts naturally have less distance between them. As a consequence, the maximum period of the correlation near-cycle must also increase. In Table 6.1, the maximum period of the correlation near-cycle is given for the BCF-3 sequence. Specifically, it is found by considering the two dimensional correlation between every pair of coordinates that satisfy the condition $\delta << 1$ in the 100 and 300 dimensional sequences. Using Figure 6.6 as a point of reference, the 100 and 300 dimensional low-discrepancy sequences are the necessary sizes to perform the QMC particle simulation for a free molecular duct with a height to length ratio L approximately equal

Sequence	Correlation Pattern	Near-period	Max Near-period $s = 100$	Max Near-period $s = 300$
Halton	$\lvert p_1 - p_2 \rvert \ll p_1, p_2$	$\frac{p_1 p_2}{\lvert p_1 - p_2 \rvert}$	$1.4 \cdot 10^5$	$1.9 \cdot 10^6$
BCF-3	$\big\lvert [z_1] - [z_2] \big\rvert \ll 1$	$\big\lvert [z_1] - [z_2] \big\rvert^{-1}$	$1.1 \cdot 10^5$	$4.1 \cdot 10^6$
Faure base q	–	q^2	$1.0 \cdot 10^4$	$9.4 \cdot 10^4$
Niederreiter[a] base 2	$w_{1,1} \sim w_{2,1}$	$2^{\alpha+1}$	$1.3 \cdot 10^5$	$2.1 \cdot 10^6$

[a]Here α is the number of leading order bits that are identical between the two 32-bit computer words, $w_{1,1}$ and $w_{2,1}$, used to generate the most significant bit of the two dimensions of the Niederreiter sequence in base 2. If $w_{1,1}$ and $w_{2,1}$ are represented as unsigned integers, *i.e.* $0 \leq w_{1,1}, w_{2,1} < 2^{32}$, then $\alpha = 31 - \lfloor \log_2(w_{1,1} \oplus w_{2,1}) \rfloor$, where $\oplus$ denotes the bit-wise XOR operation.

Table 6.1: The period of the near-cyclic construction patterns illustrated in Figure 6.17. The longest near-period is found for each example assuming the low-discrepancy sequence has 100 and 300 dimensions.

to 4 and 10, respectively. As expected, when the dimension of the sequence increases from 100 to 300, the period of the correlation near-cycle in the BCF-3 sequence increases nearly 40-fold from $1.1 \cdot 10^5$ to $4.1 \cdot 10^6$.

The length of the correlation near-cycle found for the BCF-3 sequence may be somewhat surprising, especially if one expects the fractional parts of the irrational numbers used to construct the sequence to be more or less uniformly distributed in the unit interval. If the fractional parts were in fact uniformly distributed, the resulting period of the correlation near-cycle would be orders of magnitude smaller than in Table 6.1. This is not the case because the BCF-3 sequence is specially constructed from irrational numbers represented by periodic continued fractions that only contain the coefficients 1, 2, and 3. Recall from Section 4.1 (see also [75, 131]) that the real numbers in the interval $[0, 1)$ have continued fraction representations, both finite and infinite, which may contain any positive integer as a coefficient. Thus, the possible values for the fractional parts of the irrational numbers used to construct the BCF-3 sequence are restricted to a very small subset of the unit

interval. Consequently, the fractional parts tend to be much more closely spaced when compared to a more uniform distribution among all the real numbers in the interval $[0, 1)$. This is most noticeable when the continued fraction representations of the irrational numbers are very similar. To illustrate this particular case, consider the following irrational numbers that appear in the construction of the 300 dimensional BCF-3 sequence:

$$\langle \overline{3} \rangle = 3.302775637\ldots \tag{6.9}$$

$$\langle \overline{2,3,3,3} \rangle = 2.302677894\ldots$$

$$\langle \overline{2,3,3,3,3} \rangle = 2.302784597\ldots$$

$$\langle \overline{2,3,3,3,3,3} \rangle = 2.302774816\ldots$$

$$\langle \overline{1,3,3,3,3,3,3} \rangle = 1.302775882\ldots. \tag{6.10}$$

Note that the difference δ between the fractional parts in (6.9) and (6.10) is approximately equal to $2.4 \cdot 10^{-7}$, which yields the maximum near-period given in Table 6.1 for the BCF-3 sequence.

The construction pattern illustrated in Figure 6.15(a) is not the only potential source of significant two dimensional correlation in the BCF-3 sequence. Similar to the Halton sequence using a pair of Sophie Germain prime bases, it is possible for the initial elements of the BCF-3 sequence to be restricted to a band parallel to the line $x_2 = \frac{1}{2}x_1 \pmod 1$ instead of $x_2 = x_1$. This type of construction pattern occurs between two dimensions of the BCF-3 sequence that are generated from irrational numbers z_1 and z_2 with the property $[z_2] \approx \frac{1}{2}[z_1]$. In order to expand this idea to other construction patterns, define the generalized difference function $\delta(a, b)$ for the irrational numbers z_1 and z_2 such that

$$\delta(a, b) = |az_1 - bz_2 - \overline{n}|, \tag{6.11}$$

where a and b are non-zero integers, and $\overline{n}$ is the nearest integer to $az_1 - bz_2$. Note that $\delta(1,1)$ is equivalent to the difference in the fractional parts defined earlier in (6.8). If the generalized difference $\delta(a,b)$ is very small, it implies that the irrational number z_1 is closely approximated by a rational multiple of z_2; and vice-versa. Even though, z_1 and z_2 are linearly independent over the rationals, when $\delta(a,b) \ll 1$, their behavior constructing the BCF-3 sequence is effectively the same, which results in significant and persistent correlation between the dimensions. In the general case, when a and b are relatively small, and $\delta(a,b) \ll 1$, the construction pattern restricts the initial placement of the BCF-3 sequence elements to a narrow band parallel to $x_2 = \frac{a}{b}x_1 \pmod 1$. It is interesting to note that Richtmyer in [148] proves, that the error in the QMC integral approximation using any Weyl-Richtmyer sequence depends, at least in part, on the generalized difference $\delta(a,b)$ in (6.11). More specifically, if a, b, and $\delta(a,b)$ are all relatively small, then the error on the QMC integral approximation is relatively large. However, these are the same conditions that produce significant correlation between the dimensions of the BCF-3 sequence. While not rigorous, it is reassuring that the observed connection between the low-discrepancy sequence correlation and the performance loss of the QMC particle simulation is consistent with the theoretical results of Richtmyer [148].

6.4.3 Correlation between two dimensions of the Faure sequence

Unlike the other three low-discrepancy sequences, it is difficult to determine which dimension pairs of the Faure sequence are likely to have significant correlation by only considering the constructive elements of each dimension. It is possible to gain some insight into the construction patterns using the definition of a (t,s)-sequence (see [126, 127]) since the Faure sequence is classified as a $(0,s)$-sequence. From the

definition of the $(0, s)$-sequence in base q, it is known that each block of q^2 consecutive sequence elements has exactly one point in each of the elementary intervals of the form $[\frac{m}{q}, \frac{m+1}{q}) \times [\frac{n}{q}, \frac{n+1}{q})$, for $0 \leq n, m < q$. Hence, each block of q^2 elements is evenly distributed throughout $[0, 1)^2$. Based on the same reasoning as the Halton sequence, the correlation near-cycle for the Faure sequence in base q is not greater than q^2. It should be noted that this period for the correlation near-cycle is determined without specifying the sequence dimensions; and as such, the result should apply to any pair of dimensions from the Faure sequence. Although not a rigorous verification, the near-period for the two dimensional correlation of the Faure sequence is observed to be less than or equal to q^2 for all the computational experiments performed in this investigation. In comparison to the other three low-discrepancy sequence, the period of the correlation near-cycle is typically smallest for the Faure sequence. More specifically, the maximum period of the correlation near-cycle present in the 100 and 300 dimensional Faure sequences is 10 to 20 times smaller than any of the other sequences shown in Table 6.1.

Taken as a whole, each successive block of q^2 elements has approximately the same pattern as the original block; however, the order in which the individual points are added to the sequence is a permutation of the original block. As with the other low-discrepancy sequences, each successive block of the Faure sequence is also translated by a small amount to avoid any duplicate sample points, and to continue evenly distributing the sequence elements throughout the domain. While many dimension pairs of a Faure sequence in base q have a maximum correlation near-period of q^2, it is this translation amount of each block of q^2 elements that ultimately determines the magnitude and extent of the two dimensional correlation. It is difficult, however, to determine the translation amount of each successive block of q^2 elements directly

from the constructive elements of each dimension. Because the ordering of each successive block of q^2 elements is permuted, it is perhaps easiest to assess the impact of the translation amount on the correlation by simply plotting the blocks of the Faure sequence. Unfortunately, given the need for human oversight, the graphical approach is only practical when evaluating a few specific dimensions of a given Faure sequence.

In order to determine which dimensions of a Faure sequence produce the most significant correlation, a brute force approach must be adopted. Specifically, the minimum sequence length N_{min} (6.7), necessary for the low-discrepancy sequence to reach the same level of correlation as a random sequence, is calculated for every possible dimension pair of the Faure sequence in a given base. The most persistent two dimensional correlation in the Faure sequence is then simply found by searching for the dimension pair with the largest value of N_{min}. While an exhaustive search is almost never the first choice for a numerical method, the computational cost is reasonable[17] when the base of the Faure sequence $q \leq 167$. As an example, consider the correlation between all the dimensions of the Faure sequence in base 31, the largest value of N_{min} then occurs between the 5^{th} and 6^{th} dimensions. The initial construction pattern of this example is illustrated in Figure 6.15(b).[18] However, it should be noted that this pattern is not unique, and appears in a similar form for all pairs of consecutive dimensions. In order to understand why the correlation between the 5^{th} and 6^{th} dimensions is the most significant, the first 4 blocks of 31^2 elements are plotted in Figure 6.16(a). The amount of translation between successive blocks

[17] In this context, given the patience of the author and the relative importance of the data to the overall investigation, "reasonable" means one hour of computation time on a 3.06 GHz Intel Xeon processor.

[18] Additional graphical examples are presented by Morokoff and Caflisch in [116] for the construction patterns that yield significant correlation between the dimensions of the Faure sequence.

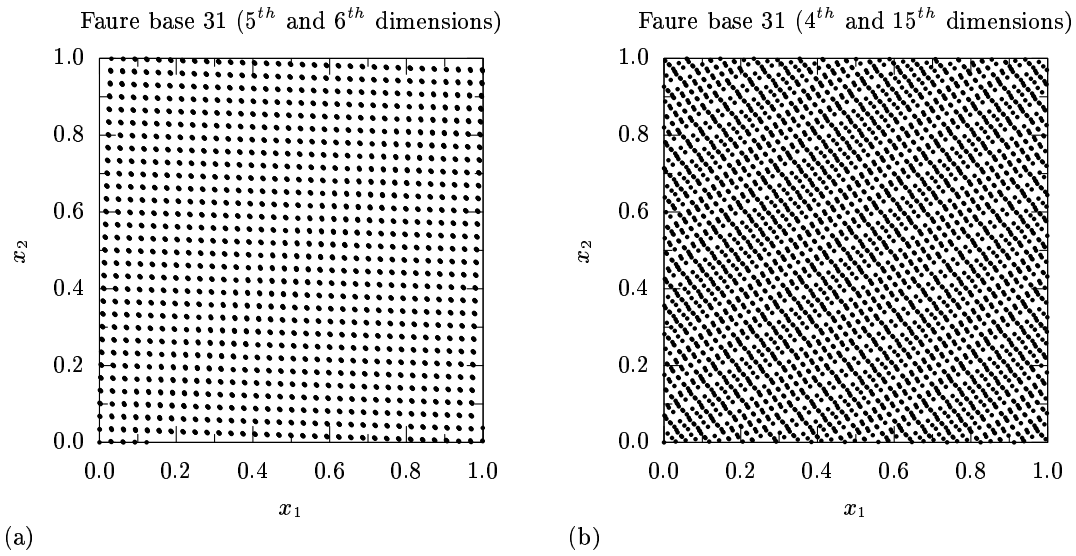

Figure 6.16: Two dimensional construction patterns that appear after the first $N = 3844$ elements of the Faure sequence in base 31: (a) dimensions 5 and 6 (largest value of N_{min}); and (b) dimensions 4 and 15 (smallest value of N_{min}).

is barely discernible with the naked eye; the slightly oval marks in Figure 6.16(a) are actually four very closely spaced circles. It is a direct consequence of this very small translation of the blocks that the correlation between the 5^{th} and 6^{th} dimensions of the Faure sequence in base 31 is the most significant. For comparison, the 4^{th} and 15^{th} dimensions of the Faure sequence in base 31 (which yield the smallest value of N_{min}), are plotted as well in Figure 6.16(b). In contrast to the 5^{th} and 6^{th} dimensions, the amount of translation between consecutive blocks of 31^2 elements is much greater, which produces a more even distribution after the same number of elements.

Given the fact that the generation of the Faure sequence is entirely deterministic, there is a direct connection between the constructive elements used to generate each dimension and the construction patterns that yield significant correlation. Although, unlike the Halton, BCF-3 and Niederreiter sequences, it is not readily apparent which dimension pairs of the Faure sequence are most likely to produce significant

correlation. The original goal of the exhaustive search for the largest N_{min} was to provide enough insight to determine this connection; however, it never became obvious to the author during the course of this investigation. The exhaustive search does provide the following observation about the correlation between the dimensions of the Faure sequence, but it does not provide any reasons as to why. For all prime bases $11 \leq q \leq 191$, the most persistent correlation (*i.e.* the largest value of N_{min}), occurs between consecutive dimensions of the Faure sequence. Unfortunately, there is not an apparent pattern between which consecutive dimensions lead to significant correlation and the base q of the Faure sequence. When the Faure base q is relatively small, there does appear to be some initial bias as to which dimension pairs produce the most significant correlation. Specifically, the largest value of N_{min} tends to occur between consecutive dimensions from either the first several or last several dimensions generated for a q-dimensional Faure sequence in base q, when $q \leq 109$. However, this initial bias is much less noticeable if all the prime bases q are considered from the range $11 \leq q \leq 191$.

6.4.4 Correlation between two dimensions of the Niederreiter sequence in base 2

For the Niederreiter sequence in base 2, it is easier to discuss the construction patterns in terms of the intervals in which the sequence elements are found rather than their actual location. In particular, significant correlation occurs between dimensions of the sequence when a large number of the initial elements are exclusively found in the shaded intervals illustrated in Figure 6.15(c). Note that any point within this union of intervals, defined by

$$\mathcal{I}_{11} = [0, \tfrac{1}{2}) \times [0, \tfrac{1}{2}) \cup [\tfrac{1}{2}, 1) \times [\tfrac{1}{2}, 1),$$

makes a positive contribution to the sum in (6.5) for the running correlation $\rho_{12}(N)$. Hence, if all the initial elements of the Niederreiter sequence in base 2 are restricted to $\mathcal{I}_{11}$, then $\rho_{12} > 0$. Since all the elements in $\mathcal{I}_{11}$ are positively correlated, the magnitude of ρ_{12} in this case is typically quite large. In fact, if the sequence elements are assumed to be uniformly distributed throughout $\mathcal{I}_{11}$, then the expected value for the running correlation is $\rho_{12} = \frac{3}{4}$.

The specific pattern for this type of construction of the Niederreiter sequence in base 2 is as follows. The first 2^α elements, where α is a positive integer to be defined later, of the sequence are more or less evenly distributed in the region $\mathcal{I}_{11}$. At this point in the construction, the Niederreiter sequence is strongly correlated with $\rho_{12}(N) \approx \frac{3}{4}$ at $N = 2^\alpha$ (based on the assumption of a uniform distribution of points). The next block of 2^α elements are similarly distributed in $\mathcal{I}_{11}\backslash[0,1)^2$; that is, the unshaded intervals shown in Figure 6.15. As a consequence of this type of construction pattern, the first $2^{\alpha+1}$ elements of the Niederreiter sequence are evenly distributed throughout the unit square; and the running correlation $\rho_{12}(N)$ is approximately zero at $N = 2^{\alpha+1}$. This pattern then continues for each successive block of $2^{\alpha+1}$ elements. Similar to the construction patterns considered for the other low-discrepancy sequences, each subsequent block of $2^{\alpha+1}$ elements is translated slightly from all the previous elements in order to continue the even distribution elements for the sequence. The only difference for the Niederreiter sequence in base 2 is that the ordering within the blocks may reverse occasionally. In particular, the first 2^α elements of each of these blocks may be restricted to either the positively correlated region $\mathcal{I}_{11}$, or the negatively correlated region $\mathcal{I}_{11}\backslash[0,1)^2$ depending on the block number. Whichever region the first 2^α elements of each block occupy, the remaining 2^α elements always occupy the opposite region. Therefore, for this type of construc-

tion pattern of the Niederreiter sequence in base 2, the period of the correlation near-cycle is $2^{\alpha+1}$.

In order to understand exactly how the Niederreiter sequence in base 2 is able to produce the construction patterns illustrated in Figure 6.15, one must return to the actual mathematical process used to generate the general sequence. Each dimension of the Niederreiter sequence in base q can be generated by the following matrix-vector multiplication

$$\mathbf{y}_n = A\vec{\xi}_q(n), \tag{6.12}$$

where all the addition and multiplication operations are defined over the finite field $\mathbb{F}_q$. The elements $a_{ij} \in \mathbb{F}_q$ of the matrix $A = [a_{ij}]$ are determined by a formal Laurent series of certain rational functions that depend on the irreducible polynomial in $\mathbb{F}_q[x]$ used to generate each dimension. The exact definition of a_{ij} can be found in Appendix F and [19, 127]. The vector $\vec{\xi}_q(n)$ is the base q representation of the integer n defined in Appendix A. Note that the least significant digits of the base q representation of n corresponds to the first elements of the vector $\vec{\xi}_q(n)$. That is, if $n = \ldots d_3 d_2 d_1 d_0|_q$ is the base q representation of n with digits $0 \leq d_0, d_1, \ldots \leq q-1$, then $\xi_{1,q}(n) = d_0$, $\xi_{2,q}(n) = d_1, \ldots$ and so forth. The vector $\mathbf{y}_n = (y_{1,n}, y_{2,n}, \ldots)$ that results from the matrix-vector multiplication in (6.12) is then used to generate the n^{th} element of this dimension of Niederreiter sequence in base q, which is denoted by x_n. Specifically,

$$x_n = \sum_{i=1}^{\infty} \frac{y_{i,n}}{q^i}. \tag{6.13}$$

The process outlined here is then repeated for each dimension of the Niederreiter sequence using the same vector $\vec{\xi}_q(n)$; however, a different irreducible polynomial in $\mathbb{F}_q[x]$ is used to produce a unique A matrix for each dimension.

Although this investigation is only concerned with the correlation between the

dimensions of the Niederreiter sequence in base 2, the following analysis may be extended to any valid base of the Niederreiter sequence. For the special base 2 case, the components of the matrix and vectors in (6.12) are in the finite field $\mathbb{F}_2$; that is, the components are either 0 or 1. Consequently, it is possible to treat these components of $\mathbf{y}_n$ and $\vec{\xi}_2(n)$ in (6.12) as individual bits and concatenate them into a single computer word for each vector. Similarly, the components of the rows (or columns) of the matrix A in (6.12) can be combined to form a single computer word for each row (or column). In addition to the reduction in the required memory, there is also a tremendous computational savings that occurs when the bits are concatenated in this manner. Details of this savings are discussed in Section 4.3 and by Bratley *et. al.* in [19]. Typically, the columns of the matrix A are treated as single computer words to improve the actual algorithm used to generate the sequence. However, for the correlation discussion here, the rows of the matrix are treated as single computer words instead. In particular, define $w_{i,j}$ as the computer word representing the components of the j^{th} row in the matrix A used to generate the i^{th} dimensions of the Niederreiter sequence in base 2. Also, define u_n as the computer word representing the components of $\vec{\xi}_2(n)$. Note that $\vec{\xi}_2(n)$ is essentially a bitwise representation of the integer n. Hence, the bits in the computer word u_n are simply reversed from the bits in the base 2 representation of the integer n; that is, the least significant bit of n is the now most significant bit of u_n.

While the process for generating the Niederreiter sequence is defined over an infinite dimensional vector space, it is necessary in practice to limit matrix and vector operations in (6.12) and 6.13) to a finite dimensional space. For the Niederreiter sequence in base 2 used in this investigation, the vector space is limited to 32 dimensions, which is the common practice. This allows for the linear system in (6.12) to be

represented with standard 32-bit computer words common to all modern programming languages. Note that this also limits the maximum sequence length that may be generated to $N < 2^{32}$. Using this representation, each component of the vector $\mathbf{y}_n$ in (6.12) can be calculated by a single bit-wise AND operation.[19] Specifically, for the i^{th} dimension of the n^{th} element of the Niederreiter sequence in base 2, the vector $\mathbf{y}_n = (y_{1,n}, \ldots, y_{32,n})$ is calculated by

$$y_{j,n} = P(w_{i,j} \otimes u_n), \tag{6.14}$$

where $\otimes$ denotes the bit-wise AND operation, and P is the parity of the argument. Let the function $\mathscr{N}_1(x)$ denote the number of bits equal to one in the base 2 representation of the integer x. The parity of a binary computer word x is then given by

$$P(x) = \mathrm{mod}(\mathscr{N}_1(x), 2).$$

That is, $P(x) = 0$ if there is an even number of 1-bits in x, and $P(x) = 1$ if there is an odd number of 1-bits in x.

Returning to the problem of correlation, consider any two dimensions of a multi-dimensional Niederreiter sequence in base 2. For simplicity, refer to these dimensions as 1 and 2, and allow the corresponding matrix and vector operations given in (6.12) and (6.14) to contain the dimension number as a superscript in parentheses. Now assume that the first rows of the matrices $A^{(1)}$ and $A^{(2)}$ used to generate their respective dimensions are identical; that is, $w_{1,1} = w_{2,1}$. In this case, the first components of the vectors $\mathbf{y}_n^{(1)}$ and $\mathbf{y}_n^{(2)}$ are the same after the matrix-vector multiplication in (6.12). Note from (6.13) that the first component of $\mathbf{y}_n$ makes the most significant contribution to the actual location of the sequence coordinate x_n; specifically, it de-

[19] Note that addition over the finite field $\mathbb{F}_2$ is equivalent to the XOR operation (exclusive-or), and multiplication over $\mathbb{F}_2$ is equivalent to the AND operation.

termines if the $x_n < \frac{1}{2}$ or $x_n \geq \frac{1}{2}$. Hence, if $y_{1,n}^{(1)} = y_{1,n}^{(2)} = 0$, then the location of the sequence element is restricted to the interval $[0, \frac{1}{2}) \times [0, \frac{1}{2})$. Conversely, if $y_{1,n}^{(1)} = y_{1,n}^{(2)} = 1$, then the location of the sequence element is restricted to the interval $[\frac{1}{2}, 1) \times [\frac{1}{2}, 1)$. Therefore, all the sequence elements are restricted to the set $\mathcal{I}_{11}$, which corresponds to the shaded region in the the construction pattern example given in Figure 6.15(c).

For most implementations of the Niederreiter sequence in base 2, the first rows $w_{1,1}$ and $w_{2,1}$ of the A matrices are never exactly the same. It is only possible for all 32 components to be the same if the irreducible polynomials used to construct the A matrices are of degree greater than 15; and polynomials this large are never used in practice. However, for any multi-dimensional Niederreiter sequence in base 2, there always exists a pair of dimensions where one or more of the most significant bits in $w_{1,1}$ and $w_{2,1}$ are the same. As a notational convenience, the relationship $a \sim b$ is used to denote that the most significant bits of the two computer words a and b are the same. Let α denote the number of the most significant bits of $w_{1,1}$ and $w_{2,1}$ that are identical. Assuming the computer words are stored as standard IEEE unsigned 32-bit integers, α can be defined mathematically by

$$\alpha = 31 - \lfloor \log_2(w_{1,1} \oplus w_{2,1}) \rfloor,$$

where $\oplus$ denotes the bit-wise XOR operation. If the most significant α bits of $w_{1,1}$ and $w_{2,1}$ are the same, then $y_{1,n}^{(1)} = y_{1,n}^{(2)}$ for all $n < 2^\alpha$; which implies that these sequence elements are restricted to $\mathcal{I}_{11}$. Note that when $n < 2^\alpha$, after excluding the most significant α bits, the remaining bits of u_n in (6.14) are all zero. Thus, any differences between $w_{1,1}$ and $w_{2,1}$ that may occur after the most significant α bits have no effect on the calculation of $y_{1,n}^{(1)}$ and $y_{1,n}^{(2)}$ when $n < 2^\alpha$.

By definition, the $(\alpha+1)^{th}$ most significant bit must be different between $w_{1,1}$ and $w_{2,1}$. Hence, $w_{1,1} \otimes u_n$ and $w_{2,1} \otimes u_n$ must have opposite parity when $2^\alpha \leq n < 2^{\alpha+1}$. If the parity is opposite, then by (6.14), $y_{1,n}^{(1)} \neq y_{1,n}^{(2)}$. This implies, that either $x_n^{(1)} < \frac{1}{2}$ and $x_n^{(2)} \geq \frac{1}{2}$, or vice-versa, when $2^\alpha \leq n < 2^{\alpha+1}$. Consequently, all the sequence elements in the range $2^\alpha \leq n < 2^{\alpha+1}$ are restricted to the set $\mathcal{I}_{11} \backslash [0,1)^2$, which corresponds to the unshaded region in the the construction pattern example given in Figure 6.15(c). The first block of 2^α elements are distributed throughout the negatively-correlated region $\mathcal{I}_{11}$, and the second block of 2^α elements are distributed throughout the negatively-correlated region $\mathcal{I}_{11} \backslash [0,1)^2$. Therefore, the period of the correlation near-cycle for this construction pattern is $2^{\alpha+1}$ as described earlier.

As the dimension of the Niederreiter sequence in base 2 increases, the maximum number of leading order bits α that are identical between any two dimensions also increases. This is true for all the computer words representing the rows of the the A matrix in (6.12), not just the first row. Without considering the actual construction of the A matrix, a lower bound on the growth of α can be established. Based on the limit of the number of unique bit combinations possible for the computer words, $\alpha \geq \lceil \log_2 s \rceil - 1$ for an s-dimensional Niederreiter sequence in base 2. However, the true growth of α for the computer words representing the first row of the A matrices (*i.e.* $w_{1,1} \sim w_{2,1}$) is much larger than this lower bound. In fact, for a 100 dimensional Niederreiter sequence, $\alpha = 16$ is the largest value found after checking every possible pair of dimensions for this construction pattern. Similarly, for a 300 dimensional Niederreiter sequence, $\alpha = 20$ is the largest value found. The maximum period of the correlation near-cycle associated with this construction pattern ($w_{1,1} \sim w_{2,1}$) is $1.3 \cdot 10^5$ and $2.1 \cdot 10^6$, for the 100 and 300 dimensional sequences, respectively. These period lengths for the Niederreiter sequence are of similar magnitude to the

correlation near-periods of the Halton and BCF-3 sequences, as indicated in Table 6.1.

It is interesting to note that the leading zeros correction for the Niederreiter sequence in base 2, suggested by Bratley *et. al.* in [19] and adopted here in this investigation, helps to reduce the value of α. The leading zeros correction modifies some of the components in the A matrix (6.12) used to generate each dimension without affecting the asymptotic convergence of the star-discrepancy of the sequence. The goal of the leading zeros correction in [19] is to help eliminate any correlation problems at the beginning of the sequence; equivalently, this reduces the value of α. If not for this correction, the number of significant bits α that are identical between the rows of the A matrices would be much higher. Consequently, there would also be a much longer period in this case for the correlation near-cycle of the Niederreiter sequence in base 2. As noted in [19], there is actually some flexibility in the implementation of the leading zeros correction. In fact, it may be possible to devise a modified leading zeros correction that is capable of further reducing the value of α. Such a strategy is very similar to the additional uniformity condition proposed by Sobol' in [162] for the direction numbers used to generate the Sobol' sequence.

Up until now, the only construction pattern considered for the Niederreiter sequence in base 2 is when $w_{1,1} \sim w_{2,1}$ (see Figure 6.15(c)). While the construction pattern for $w_{1,1} \sim w_{2,1}$ produces significant correlation between the two dimensions of the sequence, it is by no means the only known construction pattern to do so. Whenever there is a large number of significant bits that are the same between any of the first few rows of the A matrices, persistent correlation can exist between the corresponding dimensions. In order to visualize the construction pattern in this more

general case, Figure 6.15(d) shows the construction pattern that results when the 2^{nd} and 3^{rd} rows of the A matrices have the same significant bits (*i.e.* $w_{1,2} \sim w_{2,3}$). For convenience, let α_{ij} denote the number of significant bits that are identical between $w_{1,i}$ - the i^{th} row of one A matrix, and $w_{2,j}$ - the j^{th} row of the other A matrix. Similar to the $w_{1,1} \sim w_{2,1}$ case, when $w_{1,2} \sim w_{2,3}$, the first $2^{\alpha_{23}}$ sequence elements are restricted to the shaded region, and the next $2^{\alpha_{23}}$ sequence elements are restricted to the unshaded region in Figure 6.15(d). It is possible that the value of α_{ij} for the construction pattern $w_{1,i} \sim w_{2,j}$ $(1 < i, j \leq 32)$ is actually greater than the value of α_{11}, especially when the row numbers i and j are large. Hence, the resulting construction pattern for these cases will yield a longer period for the correlation near-cycle than the $w_{1,1} \sim w_{2,1}$ example given in Table 6.1.

It is important to note that as the row numbers i and j increase, the checkerboard construction pattern becomes finer, which produces a more even distribution of points as illustrated in Figure 6.15. In order to estimate the impact of this finer construction pattern on the correlation, one may assume that the sequence elements are uniformly distributed throughout the shaded regions in Figure 6.15. Under this assumption, the running correlation defined in (6.5) has an expected value of $\rho_{12} = \frac{3}{4}$ for the $w_{1,1} \sim w_{2,1}$ construction pattern; and $\rho_{12} = \frac{3}{32}$ for the $w_{1,2} \sim w_{2,3}$ construction pattern. In general, the running correlation ρ_{12} decreases when the row numbers i and j increase for the $w_{1,i} \sim w_{2,j}$ construction pattern. While the correlation near-cycle may be longer for the $w_{1,i} \sim w_{2,j}$ case when $1 < i, j \leq 32$; its overall impact may not be as significant as the $w_{1,1} \sim w_{2,1}$ construction pattern.

There is a key difference between the construction pattern of the Niederreiter sequence in base 2 when $w_{1,1} \sim w_{2,1}$, and the construction patterns of the other low-discrepancy sequences illustrated in Figures 6.12 and 6.15. The first half of

each correlation near-cycle in this construction of the Niederreiter sequence makes an exclusively positive (or negative) contribution to the sum in (6.5) for the running correlation $\rho_{12}(N)$. In contrast, the first half of the correlation near-cycles for any of the other sequences makes both positive and negative contributions to the running correlation. Thus, the $w_{1,1} \sim w_{2,1}$ construction pattern of the Niederreiter sequence in base 2 initially produces a much larger running correlation than the other sequences. As a direct consequence, many more near-cycles must be completed in order to break up this stronger form of correlation present in the Niederreiter sequence in base 2.

6.4.5 The extent of the correlation present in the low-discrepancy sequences

In order to better evaluate the impact of correlation on the QMC particle simulations, the minimum sequence length N_{min} (6.7), at which a given sequence is considered as uncorrelated as a random sequence, is calculated for each of the low-discrepancy sequences tested in this investigation. More specifically, the largest value of N_{min} is found among the first $s \leq 100$ dimensions of the low-discrepancy sequences, as shown in Figure 6.17. It is important to note that the largest value of N_{min} in Figure 6.17 is found by considering only one specific construction pattern for the Halton, BCF-3 and Niederreiter sequences. In particular, the largest value of N_{min} for the Halton sequence is found by considering dimension pairs generated from twin prime bases, an illustration of the construction pattern is found in Figure 6.12. Moreover, it is this construction pattern that consistently produces the greatest correlation between all Halton dimension pairs previously illustrated in Figure 6.14. For the BCF-3 sequence, only the construction pattern illustrated in Figure 6.15(a) is considered; that is, when the fractional parts of the irrational numbers used to

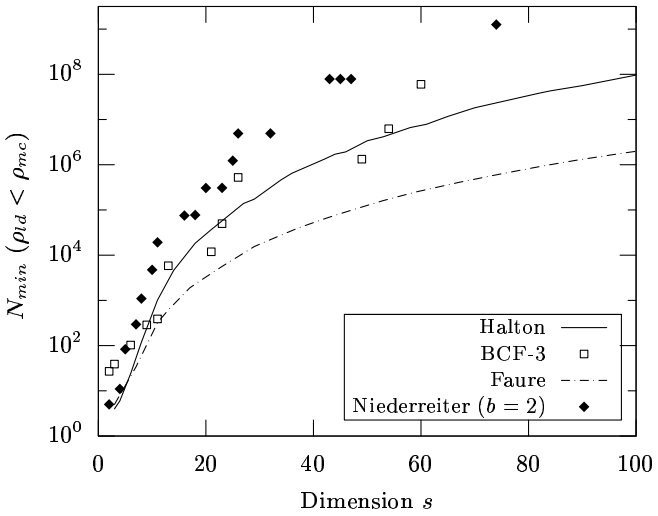

Figure 6.17: The extent of the two dimensional correlation present among the first 100 dimensions of the low-discrepancy sequences.

generate each dimension are nearly the same. Similarly, for the Niederreiter sequence in base 2, only the construction pattern illustrated in Figure 6.15(c) is considered; that is, when the words $w_{1,1}$ and $w_{2,1}$ used to generate the most significant bit of each dimension are nearly the same. As discussed earlier, it is difficult to determine which dimension pairs and construction patterns of the Faure sequence are likely to yield significant correlation. Thus, the largest value of N_{min} is found by an exhaustive search of all dimension pairs of a q-dimensional Faure sequence in base q, where the prime $q \leq 101$.

It appears from Figure 6.17 that the extent of the correlation is consistently the least between the dimensions of the Faure sequence when compared to the other low-discrepancy sequences. In particular, for all prime bases $q \leq 97$, the largest value of N_{min} found between any two dimensions of a q-dimensional Faure sequence is less than $2 \cdot 10^6$. By contrast, the extent of the correlation is typically the greatest for the Niederreiter sequence in base 2. More specifically, the largest value of $N_{min} \approx 10^9$ is found for the correlation between the first 75 dimensions of the sequence. The

fact that the Niederreiter sequence in base 2 yields the largest values of N_{min} is consistent with the earlier observations for the construction pattern. In particular, this Niederreiter sequence yields the strongest initial correlation, and many more correlation near-cycles must be completed before it is broken up. The minimum sequence lengths N_{min} of the Halton and BCF-3 sequences tend to be bounded by the other two sequences; and both show a similar increase in N_{min} with the sequence dimension s. In fact, the Halton and BCF-3 sequences both attain a maximum value of $N_{min} \approx 10^8$ within the first 100 dimensions of the respective sequence, although this maximum occurs much earlier for the BCF-3 sequence ($s = 60$) as compared to the Halton sequence ($s = 99$).

The extent of the correlation present between the dimension pairs of the low-discrepancy sequences in Figure 6.17 does introduce some doubt into the physical accuracy of the QMC particle simulations. However, in spite of this correlation, the actual results in Section 6.3 for the QMC particle simulation clearly show significant gains in both speed and accuracy when the sequence dimension $s < 100$. As Morokoff and Caflisch note in [116], the mere presence of correlation in a low-discrepancy sequence does not prevent a QMC method using the sequence from producing an accurate approximation. The amount of accuracy lost by the presence of correlation in a low-discrepancy sequence used by a QMC method ultimately depends on the physical problem being simulated. As an extreme example, consider the QMC integration of a function $f(\mathbf{x})$, with $\mathbf{x} \in \overline{I}^s$, defined as a simple sum of the coordinates; that is, $f(\mathbf{x}) = x_1 + \cdots + x_s$. Since the integral of $f(\mathbf{x})$ is equivalent to the sum of s one dimensional integrals, any correlation between the dimensions of the low-discrepancy sequence used in the QMC method has no effect on the accuracy if each dimension is well-distributed in $[0, 1)$. Conversely, if there is substantial

inter-dependence among the dimensions of the function being integrated, then any correlation in the low-discrepancy sequence is likely to produce a physically inaccurate QMC approximation. It is important to remember that the Koksma-Hlawka inequality ensures that correlation between the dimensions of the low-discrepancy sequences, and the corresponding problems with physical accuracy, do not last forever. Once the sequence length is sufficiently long, all the dimension pairs of a low-discrepancy sequence become as uncorrelated as a random sequence. However, this sequence length may be too long to simulate in practice.

For the QMC simulation of the free molecular conductance probability, significant correlation in the low-discrepancy sequence typically occurs between the higher dimensions. The higher dimensions of the physical problem correspond to the wall collisions that occur when the particle weight is lower; and hence, the impact on the sample trajectory score is smaller. For a free molecular duct geometry $L \leq 4$ (*i.e.* problem dimension $s < 100$), there are only a few dimension pairs among the low-discrepancy sequences in Figure 6.17 that have a value N_{min} greater than the number of samples collected for the QMC simulation. Moreover, these few correlated dimension pairs correspond to particle collisions when the particle weight is near the truncation weight. It is not surprising then to see that the QMC particle simulations are quite accurate when $L \leq 4$. Unfortunately, for a free molecular duct geometry $L = 10$ (with a problem dimension $s \approx 300$), there are many dimension pairs that remain significantly correlated well beyond the number of samples collected for the QMC simulation. Worse yet, the weight of the simulated particle at these correlated dimensions is typically greater than 1% of the initial weight. Hence, the QMC simulation of the duct geometry $L = 10$ contains a non-physical representation of the diffuse wall collision process, which explains, in part, the poor performance of the

method in this case.

It is important to remember that the presence of correlation alone is not enough to condemn a QMC simulation. The correlation analysis introduced here is best used as a tool to check if there is any physically inconsistent behavior in the simulation, and to locate where within a simulation it may occur. In this context, the following are the two key points of the correlation results presented in this investigation. In order to be certain that a QMC method is physically consistent with the problem being simulated, one should collect a sufficient number of samples to ensure the correlation present between any dimensions of the low-discrepancy sequence is negligible. If this is not feasible because the necessary number of samples is intractably large, then it is still possible to obtain an accurate approximation if the simulation is designed in such a manner as to ensure the correlated dimensions have little impact on the physical problem.

6.5 Hybrid Quasi-Monte Carlo Simulation

Hybrid quasi-Monte Carlo and Monte Carlo (QMC/MC) integration refers to any method that uses both techniques to approximate an integral, the aim of which is to produce a composite method that retains the positive features of both approximations while avoiding their negative aspects as much as possible. As illustrated in Sections 6.3 and 6.4, the QMC-only particle simulation suffers a decrease in the error convergence rate as the duct length to height ratio L increases. This performance loss is attributed to an increase in the non-physical correlation between the molecular moves that occurs when the dimension of the low-discrepancy sequence used in the QMC simulation increases. Ultimately, there is a practical upper limit on the dimension of the low-discrepancy sequence that can be used to obtain a computationally

efficient QMC approximation, which depends on the physical problem being simulated. The low-discrepancy sequence length limitation present in the QMC method is one of the problems the hybrid QMC/MC method is intended to ameliorate by reducing the dimension of the low-discrepancy sequence needed for the QMC portion of the simulation.

Borrowing the terminology from Spanier in [167], there are two common strategies for implementing this type of hybrid QMC/MC simulation: the *mixed* strategy, and the *scrambled quasirandom* strategy. The mixed strategy simply replaces certain dimensions of the low-discrepancy sequence used in the original QMC-only simulation with pseudo-random numbers, thus reducing the dimension of the low-discrepancy sequence needed. As an example, consider the mixed strategy for a problem with d physical dimensions using a s dimension low-discrepancy sequence where $s < d$. Each sample for the integral approximation in this case requires a vector $\mathbf{x} = (x_1, \ldots, x_d) \in \overline{I}^d$, which is generated by the two methods: the s elements $\{x_{i_1}, \ldots, x_{i_s}\}$ are generated by the low-discrepancy sequence, and the remaining $(d - s)$ elements $\{x_{i_{s+1}}, \ldots, x_{i_d}\}$ are generated by a pseudo-random sequence. Note that $i_1, \ldots, i_d$ represent the distinct indices $1, \ldots, d$, and the low-discrepancy and pseudo-random sequences can be applied to any ordering of the sample dimensions. This type of strategy is especially effective for simulations where the relative impact on the final solution is known for each sample dimension in the physical problem. In such a case, one restricts the application of the low-discrepancy sequence to the dimensions of the problem that most dominate the final solution. Physical problems that can be represented as integral equation with an absolutely convergent Neumann series solution, such as the conductance probability (5.27), are one class of problems where the relative impact of each dimension is known *a priori*. The

mixed strategy for the hybrid QMC/MC method is investigated by Spanier [167] and Spanier and Li [168], wherein they apply the method to several model transport problems.

The scrambled quasirandom strategy for implementing a hybrid QMC/MC simulation is not as closely linked to the physical problem as the mixed strategy. The mixed strategy reduces the dimension of the low-discrepancy sequence needed for the simulation by generating pseudo-random numbers for some of the sample dimensions. In contrast, the scrambled quasirandom strategy employs a low-discrepancy sequence with fewer dimensions than the physical problem and reuses each low-discrepancy sequence dimension multiple times to generate all the dimensions of the problem. A word of caution is in order because repeated use of the same dimension of a low-discrepancy sequence to generate independent events within the same sample can have disastrous effects on the accuracy of the simulation. In the case of the van der Corput simulation of the conductance probability given in Section 5.5 (see Figures 5.9 and 5.10), the correlation inherent in the low-discrepancy sequence construction leads to particle behavior that is not physically consistent with the actual problem. Thus, great care must be exercised when selecting the order to reuse the dimensions of the low-discrepancy sequence. Problems with non-physical correlation are avoided in the scrambled quasirandom strategy by randomly permuting a subsequence of each dimension reused in the low-discrepancy sequence.

To demonstrate, suppose one wanted to generate N samples for a problem with $d = ks$ physical dimensions using a s dimension low-discrepancy sequence. The scrambled quasirandom strategy requires a low-discrepancy sequence with a length kN to accomplish this. The first N members of this low-discrepancy sequence (recall that each member is a $s-$tuple) are used to generate the first s dimensions of each of

the N samples. The second N members of the low-discrepancy sequence are then used to generate the second s dimensions of each sample. The second N members of the low-discrepancy sequence, however, cannot be applied to the same order of samples as the first N members without introducing non-physical correlation into the simulation. To avoid this problem, the scrambled quasirandom strategy randomly permutes the order in which the second N members are applied to the second s dimensions of the samples. Each remaining low-discrepancy subsequence of N members is randomly permuted in the same manner to generate the remaining dimensions of the N samples. A more thorough description of the scrambled quasirandom strategy for the hybrid QMC/MC methods is given by Spanier in [167], where the method is applied to model transport problems. The scrambled quasirandom strategy has been developed for several other applications, including the Krook and Wu solution to the Boltzmann equation (Lécot in [88]); the heat equation (Morokoff and Caflisch in [115]); and the diffusion of quantum mechanical systems (Moskowitz in [119]).

There are, unfortunately, two drawbacks to the scrambled quasirandom strategy that are not present in either the Monte Carlo, quasi-Monte Carlo, or mixed QMC/MC methods. First, the number of samples N must be selected *a priori*. As noted in Chapter III, one of the appealing aspects of Monte Carlo and QMC is the ability to continually add new samples to improve the integral approximation without wasting previous calculations or needing to add a significant number of new samples to reach the next level of refinement. It is possible to refine the scrambled quasirandom strategy without wasting the first N samples; however, it must be done by adding blocks of N new samples. Second, there is an increase in the memory storage requirements. One must either store the information about the N samples being generated, or one must pre-compute and store the N members of the low-

discrepancy subsequence being randomly permuted. It is possible to avoid the extra storage by using a linear congruential pseudo-random generator [78] to determine the permutations in advance, which allows an individual sample to be constructed in its entirety, rather than in blocks of d dimensions. However, the non-sequential construction of the low-discrepancy sequence is far more costly than generating the low-discrepancy sequence in its natural order.

The mixed strategy and the scrambled quasirandom strategy represent hybrid QMC/MC simulations that seek to improve on the performance of a QMC-only simulation by reducing the dimension of the low-discrepancy sequence. It is precisely this type of performance gain that is of interest to this investigation. There is, however, another application of the hybrid QMC/MC found in the literature that has an alternative goal. Rather than reduce the number of dimensions needed for the low-discrepancy sequence, the hybrid QMC/MC integration is used to improve the error estimation of the QMC method. One of the nice aspects of the QMC method is that the Koksma-Hlawka inequality (3.5) provides a deterministic upper bound to the integration error. Unfortunately, as discussed by Morokoff [116] and noted in Chapter III, the Koksma-Hlawka is not generally a tight bound on the error when the number of samples is small or the problem dimension is large. This alternate approach to hybrid QMC/MC integration randomizes the actual construction of the low-discrepancy sequence rather than its implementation or application to a physical problem. The resulting hybrid QMC/MC method then allows for probabilistic estimates of the integration error in a manner similar to the traditional Monte Carlo method. As a specific example of this approach, Cranley and Patterson [35] randomly shift the origin of the integration lattice (modulo 1) to produce a hybrid QMC/MC integration technique using Korobov's method of good lattice points (see

Section 3.4). More recently, Owen [135] constructs a (t, s) sequence in base b where the digits $(0, 1, \ldots, b-1)$ used in the representation of each sequence element are randomized through a specially defined permutation.

In this investigation, the mixed strategy is selected for the hybrid QMC/MC method because it offers the easiest implementation and a direct connection to the physical problem being simulated. As an added benefit, the sample dimensions with the greatest impact on the final solution are already known. The QMC-only simulation developed in Section 5.5 is based on the absorption weighted technique for variance reduction, which prevents the simulated test particles from completely escaping the interior of the duct. Instead of the particles directly escaping, the probability of a particle escaping the duct is calculated at each location the particle intersects with the wall. The weight of the simulated test particle is reduced after each move by a factor equal to this escape probability, and the probability the test particle escapes through the outlet on its next move is then tallied for the conductance probability. For each sample trajectory generated by the absorption weighted technique, the weight of the simulated test particle monotonically decreases after each particle move. As the weight of the test particle decreases, so does its impact on the tally collected for the conductance probability. It is best to use the low-discrepancy sequence to generate the first moves of the test particle trajectory, when adopting the mixed strategy for the hybrid QMC/MC simulation.

Applying the low-discrepancy sequence to the first particle moves essentially divides the simulation into two different physical problems. Henceforth, let s denote the length of the low-discrepancy sequence used in the hybrid QMC/MC simulation proposed here. The hybrid QMC/MC method calculates the probability a particle escapes the outlet Ψ_{qmc} within the first s moves as a QMC simulation, and the prob-

ability a particle escapes the outlet Ψ_{mc} in more than s moves as a Monte Carlo simulation. The combination of these two simulations provides an approximation to the conductance probability $\Psi \approx \Psi_{qmc} + \Psi_{mc}$, with $\Psi_{qmc} \gg \Psi_{mc}$. The only interaction, or communication, between these two simulations is that the QMC method provides the position of the particle after s moves as the initial position for the Monte Carlo simulation.

In the mixed strategy described by Spanier in [167], the pseudo-random sequence is applied to the same formulation of the integral approximation as the low-discrepancy sequence. An equivalent mixed strategy for the hybrid QMC/MC simulation of the conductance probability would thus generate Ψ_{mc} using the absorption weighted Monte Carlo technique. The work ratio indicates the traditional DSMC method is faster at reaching a given error than the absorption weighted Monte Carlo technique, for most duct geometries under consideration (see Figure 5.13(b)). Further support of the computational advantage held by the traditional DSMC method method over the absorption weighted Monte Carlo technique is found in Figure 6.2. While the calculations of Ψ and Ψ_{mc} are based on the same stochastic movement of the particles within the duct, there are slight differences found in the initial particle weight and distribution. These differences, however, affect both the DSMC method and the absorption weighted Monte Carlo technique in the same manner, and are effectively canceled out in their variance ratio. To obtain the most computationally efficient approach to the calculation of Ψ_{mc}, and thus the calculation of the conductance probability Ψ, the traditional DSMC test particle method is used instead of the absorption weighted Monte Carlo technique. With this modification to the mixed strategy described by Spanier, the hybrid QMC/MC simulation proposed here uses the QMC method for the first s particle moves (Ψ_{qmc}), and the DSMC test particle

method for the remaining particle moves (Ψ_{mc}).

The hybrid QMC/MC method is used in this investigation to simulate the conductance probability Ψ in two different duct geometries ($L = 2$ and $L = 10$) using the three fastest low-discrepancy sequences for the QMC portion of the method: the Halton sequence; the BCF-3 sequence; and the Niederreiter sequence in base $b = 2$. The QMC-only simulation of the $L = 2$ case requires a low-discrepancy sequence with dimension $s = 44$. In comparison, the hybrid QMC/MC simulation for the $L = 2$ case is tested for low-discrepancy sequences with dimensions $s = 4$, 8, and 16. The QMC-only simulation of the $L = 10$ case requires a low-discrepancy sequence with dimension $s = 306$, and the hybrid QMC/MC simulation is tested for low-discrepancy sequences with dimensions $s = 32$, 64, and 128. In Figure 6.18, the convergence of the relative error is found for these hybrid QMC/MC simulations and compared to the QMC-only, DSMC, and absorption weighted Monte Carlo methods. The hybrid QMC/MC simulation does not improve the error convergence of the QMC-only simulation, in any of the cases tested in this section. However, the error convergence of the hybrid QMC/MC simulation using a low-discrepancy sequence with $s = 16$ for the $L = 2$ case is nearly the same as the QMC-only simulation. Similarly, the error convergence of the hybrid QMC/MC simulation using a low-discrepancy sequence with $s = 64$ and 128 for the $L = 10$ case is also the same as the QMC-only simulation. Therefore, with little or no sacrifice to the accuracy of the simulation, the hybrid QMC/MC method is able to achieve the same results as the QMC-only simulation with 3 to 10 times fewer low-discrepancy sequence dimensions.

The reduction of the low-discrepancy sequence dimension used in the QMC portion of the hybrid method yields a lower overall computation time for the method. A comparison of the computation time of the hybrid QMC/MC simulation and the

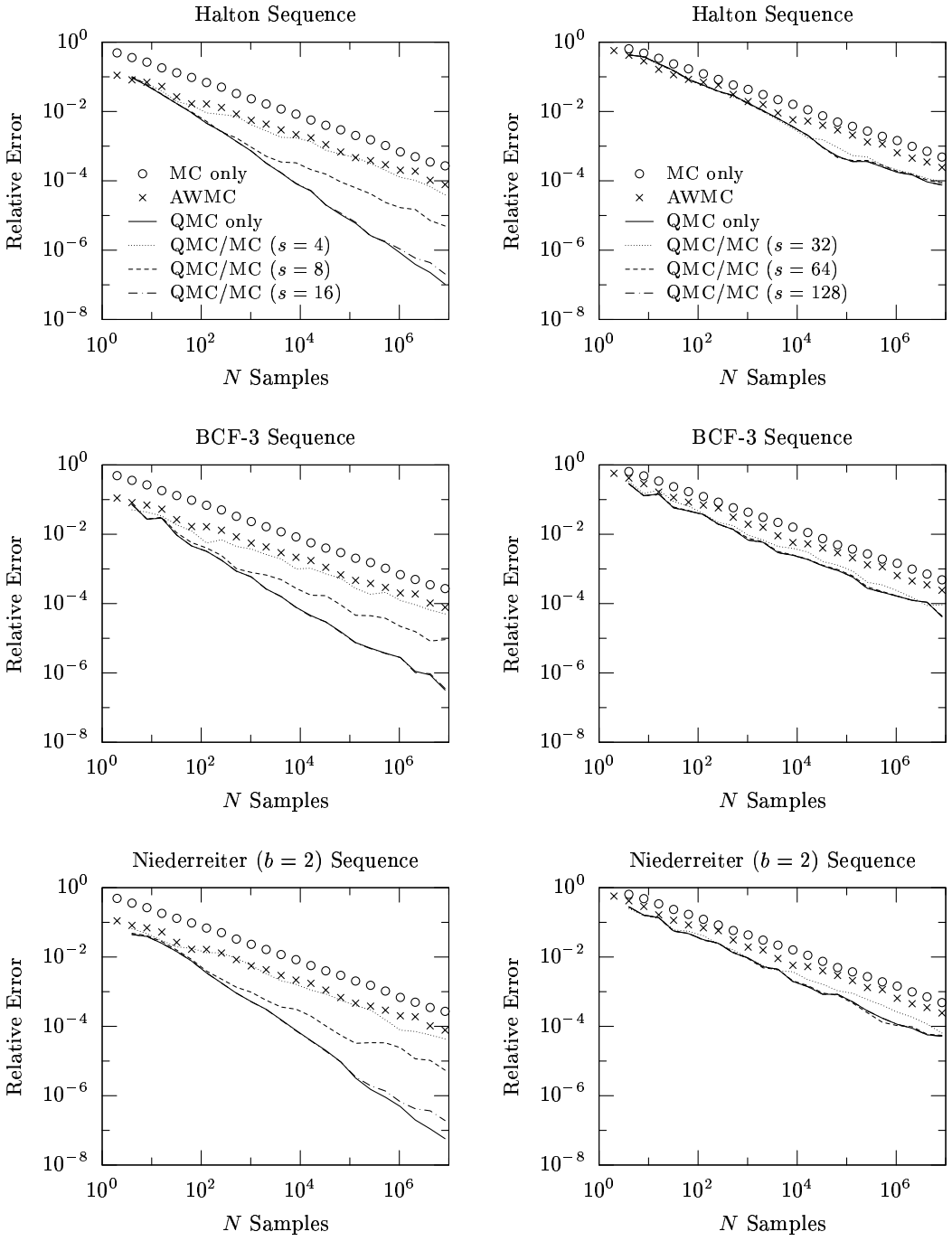

Figure 6.18: Convergence of the relative error for the hybrid QMC/MC simulation using different low-discrepancy sequences: duct length to height ratio $L = 2$ (left column); and duct length to height ratio $L = 10$ (right column).

		Computation Time[a]		
L	s	Halton	BCF-3	Niederreiter[b]
2	4	5.73	2.66	2.49
2	8	7.95	4.84	4.43
2	16	12.0	9.22	8.27
2	44[c]	25.1	24.3	21.3
10	32	8.04	7.46	6.69
10	64	13.5	13.7	12.2
10	128	24.4	26.3	23.1
10	306[c]	53.6	60.1	52.2

[a]The computation time for each simulation is normalized by the time required to generate an equivalent number of sample trajectories using the Monte Carlo method.

[b]Niederreiter base 2 sequence.

[c]QMC-only simulation.

Table 6.2: Comparison of the computation times of the hybrid QMC/MC simulation for different low-discrepancy sequences(for $L = 2$ and 10).

QMC-only simulation is given in Table 6.2. Note that the computation time presented in Table 6.2 is for the same number of samples generated, not for reaching the same error level. The timing results for the hybrid QMC/MC simulation indicate that the same error level can be reached 2-2.5 times faster for the $L = 2$ case, and 4-4.5 times faster for the $L = 10$ case than the QMC-only simulation. Furthermore, the results in Table 6.2 for the hybrid QMC/MC simulation effectively increase the error levels (at which $\tau_{qmc} < \tau_{mc}$) that appear in Figure 6.10. The hybrid QMC/MC simulation proposed here is shown to be computationally faster than the traditional test particle Monte Carlo simulation over a wider range of duct geometries and error levels than the QMC-only simulation.

The hybrid QMC/MC simulations tested here demonstrate the potential performance improvement available and are not necessarily the best combination of the QMC and Monte Carlo methods. The cases presented in Figure 6.18 and Table 6.2 represent the general trade-off between computation time and accuracy of the hybrid

method with no attempt made to optimize the performance gains. It is possible to tune the performance of the hybrid QMC/MC simulation by simply changing the dimension of the low-discrepancy sequence used in the QMC portion of the simulation. It may also be possible to achieve some performance gains by adjusting the accuracy of the Monte Carlo portion of the simulation through oversampling. In essence, the particle that remains in the duct after the QMC simulation can be split a small number of times, and the Monte Carlo simulation is then applied separately to each fraction of the particle. While it is of interest to understand the best combination of the QMC and Monte Carlo methods for developing faster particle simulations, such a study does not appear in this investigation and is reserved for future research.

CHAPTER VII

CONCLUSIONS

Accurate and efficient simulation of low-speed non-equilibrium gas flows has remained a much sought, yet elusive, goal in fluidic Micro-Electro-Mechanical-Systems (MEMS) research. The two most popular simulation techniques for MEMS applications involving gas flows are the Navier-Stokes solution with slip boundary conditions and the direct simulation Monte Carlo (DSMC) method of Bird. In almost every application, the Navier-Stokes solution has a much lower computational cost than the DSMC method. Unfortunately, it is only physically accurate for near-equilibrium gas flows; that is, when the Knudsen number $Kn \lessapprox 0.1$. The DSMC method, in contrast, is physically valid for the entire range of Knudsen numbers $0 < Kn < \infty$. However, the DSMC method has a substantially larger computation cost due to its relatively slow convergence rate $\mathcal{O}\left(N^{-1/2}\right)$, where N is the number of samples. If the average bulk velocity is significantly slower than the average speed of the simulated particles, which is common in many fluidic MEMS, then the problems associated with the slow convergence are exacerbated. In fact, when the average velocity in the fluidic MEMS is on the order of [mm/sec], the computation time of the DSMC method is often intractably long on all but the world's largest supercomputers.

In light of these challenges facing the current state of the art, two approaches

were explored in this investigation to improve the simulation methods for low-speed non-equilibrium gas flows. The first approach (see Chapter II) was to design and test empirical corrections to the computationally efficient Navier-Stokes solution in an effort to obtain greater accuracy when $Kn > 0.1$. The second, and more ambitious, approach (see Chapters III-VI) was to develop a quasi-Monte Carlo (QMC) particle simulation that retains the physical accuracy of the DSMC method, while achieving a nearly linear error convergence rate of $\mathcal{O}\left(N^{-1+\epsilon}\right)$ (for all $\epsilon > 0$), which is superior to traditional Monte Carlo methods. At the conclusion of this investigation, neither approach was considered as a general replacement to the current simulation techniques for low-speed non-equilibrium gas flows. There was, however, specific cases of practical interest in which the two new approaches yielded noticeable improvement over the current state of the art. Furthermore, the insight gained from development of these new approaches will serve to guide future designs of an accurate and efficient simulation of low-speed non-equilibrium gas flows. In this concluding chapter, the main results of this investigation are highlighted in Section 7.1; and a few notable implications of these results for further research are presented in Section 7.2.

7.1 Summary

Empirical corrections to the Navier-Stokes equations were evaluated in Chapter IV for Couette and Poiseuille flows in the transition regime ($0.01 \leq Kn \leq 10$) concerning the first approach toward an accurate and efficient simulation. These empirical corrections included: (i) a velocity slip coefficient C_s for the boundary conditions; and (ii) a viscosity correction C_μ for the shear stress closure. Empirical, or unified, models that correct both the boundary conditions and the transport closures have previously appeared in the literature for the transition regime (*e.g.* the KB

model of Karniadakis and Beskok [15], and the BPB model of Bahukudumbi, Park and Beskok [12]). However, these empirical models do not address two important questions, which should be answered if this technique is to be applied to general low-speed non-equilibrium gas flows. First, the actual construction details of these models are somewhat vague, thus prompting the question, "what, if any, physical validity exists in the empirical corrections to the Navier-Stokes equations in the transition regime?" Second, these models were only tested on a narrow range of flow conditions for which they were specifically designed, leaving one to wonder, "what are the actual predictive capabilities of such a scheme?" To address these questions, new empirical models for C_s and C_μ were constructed in Chapter II of this investigation.

The corrections C_s and C_μ for the new empirical models were determined from the Navier-Stokes solution that best fits, in a linear least squares sense, a set of known non-equilibrium solutions calculated by the DSMC method of Bird. Similar to the KB and BPB models, the new empirical models were designed to capture the Knudsen number dependence of the corrections $C_s(Kn)$ and $C_\mu(Kn)$ through the use of non-linear model laws. Further, a new feature was introduced to the model design of this investigation; specifically, the sensitivity of the corrected Navier-Stokes solution to $C_s(Kn)$ and $C_\mu(Kn)$ was included in the Levenberg-Marquardt non-linear curve-fitting method. By ensuring that the empirical models best fit the known non-equilibrium results when the solution sensitivity was the greatest, the corrected Navier-Stokes solution was shown to yield a more uniform accuracy throughout the transition regime. The drawback to the new empirical models developed in this investigation (as well as the KB and BPB models) is that they are constructed from a database of known non-equilibrium solutions, which must be provided somehow

by a more accurate method.[1] It is therefore difficult to separate the extent to which the empirical models are actually predicting a non-equilibrium solution versus simply interpolating (or extrapolating) from the database of known solutions. To address the physical validity and predictive power of the new empirical models, several test cases outside the non-equilibrium database were simulated in Chapter II, which included: (i) interpolation and extrapolation of the database of non-equilibrium solutions for Couette and Poiseuille flow; (ii) combined Couette and Poiseuille flow; (iii) variations in the tangential momentum accommodation coefficient (TMAC); (iv) different gas species; and (v) Poiseuille flow with uniform suction and injection.

Based on the results of these test cases, the following conclusions were made regarding the empirical corrections to the Navier-Stokes solution in the transition regime. The choice of empirical model was largely irrelevant in the near-equilibrium regime (*i.e.* $Kn \lessapprox 0.1$) because slip model correction is mathematically consistent with the Boltzmann equation in the continuum limit $Kn \to 0$. As the Knudsen number increased, the physical validity and predictive power of the corrected Navier-Stokes solution diminished. The method was able to capture small deviations from the non-equilibrium database; for example, the interpolation of the database flow conditions and combined Couette and Poiseuille flows. Any changes in the physical processes that occurred at the molecular level, however, were not accurately represented by the corrected Navier-Stokes solution. These changes included: the wall accommodation, and the molecular weight of the gas species. The accuracy of the empirical corrections was also extremely sensitive to the flow geometry in the transition regime, yielding grossly inaccurate predictions when the Poiseuille model was

[1]The new empirical models proposed in this investigation are based on a database of DSMC solutions for Couette and Poiseuille flow while the KB and BPB models are based on a database of solutions to the one dimensional linearized Boltzmann equation (see Sone *et. al.* [165]).

applied to Couette flow (or vice-versa) when $Kn \gtrapprox 0.1$. This investigation therefore recommends that one should limit the use of the empirical correction to the Navier-Stokes solution in the transition regime to cases when many different scales of the same geometry must be evaluated, and/or a crude estimate of the flow properties is needed.

For the second approach to an efficient and accurate simulation of low-speed non-equilibrium flow, a QMC particle simulation was developed based on the ideas of Chapter III and Chapter IV. The basic theory regarding the convergence of the QMC method was reviewed in Chapter III; in particular, the Koksma-Hlawka inequality was discussed with a focus on the concepts of variation and discrepancy. The QMC method developed in this investigation was implemented with the following low-discrepancy sequences: (i) the Weyl-Richtmyer sequence; (ii) the Halton sequence; (iii) the Faure sequence; and (iv) the Niederreiter sequence in base 2. The actual design of the algorithms used to generate these sequences is given in Chapter IV.

In addition to this algorithm review, a new construction of the Weyl-Richtmyer sequence was proposed in Section 4.1 based on heuristic arguments which suggested that the sequence would be well-suited for certain types of QMC particle simulations. This low-discrepancy sequence, termed the BCF-3 sequence, was then found to have the lowest empirical bounding constant on the discrepancy of each dimension when compared to other Weyl-Richtmyer sequences that appear in the literature, which was the intended design. More importantly, the QMC particle simulations using the BCF-3 sequence possessed the smallest power law constant in the empirical models of the error convergence data. A smaller power law constant for the error convergence data indicated that the initial accuracy of the simulation was greater after a relatively small number of samples. The QMC particle simulations using

the BCF-3 sequence was thus faster than the traditional test particle Monte Carlo method for the largest range of simulation accuracies when compared to the other Weyl-Richtmyer sequences.

A QMC particle simulation was developed in Chapter V to calculate the conductance probability for free molecular flow in a two dimensional duct. Other solution techniques for this flow were also reviewed in an effort to better understand the physical processes being simulated and to validate the performance gains achieved by the new QMC particle simulation. These included: (i) a Markov chain simulation; (ii) a finite-state linear system solution; (iii) the Nyström method using Gauss-Legendre quadrature; and (iv) the traditional test particle Monte Carlo method. In order to guide future development of particle simulations, two well-meaning, but unsuccessful, attempts at the QMC method were presented in Section 5.5, along with the reasons for their failure. As a consequence of these failed schemes, two important lessons were demonstrated in this investigation. First, a physically accurate QMC particle simulation was not obtained simply by the replacement of the pseudo-random number generator in the DSMC method with a one dimensional low-discrepancy sequence. The elements of each dimension of any low-discrepancy sequence in general are highly dependent on one another, and thus they are not a physically valid representation of a sequence of random variates. Second, it is important that all the discontinuous YES/NO decisions were eliminated in the QMC particle simulation. The presence of such discontinuities indicated the integrand for the QMC method was not of bounded variation in sense of Hardy and Krause, which prevents the Koksma-Hlawka inequality from establishing a theoretical error bound on the method. Furthermore, the observed error convergence rate was only slightly faster than the Monte Carlo method $\mathcal{O}\left(N^{-1/2}\right)$, which was still substantially slower than

the theoretical near-linear limit. By incorporating these two lessons, the final version of the QMC particle simulation developed in Section 5.5 was shown to achieve a near-linear error convergence rate.

The new QMC particle simulation for free molecular flow was then tested in Chapter VI for 20 different duct geometries with a length to height ratio $0.5 \leq L \leq 10$. The QMC particle simulation clearly demonstrated near-linear error convergence for the wider duct geometries (*i.e.* smaller values of L), which led to significant performance gains over the test particle Monte Carlo method. For duct geometries with $L \leq 5$, the accuracy of the QMC particle simulation was between 90 ($L = 5$) and 33,000 ($L = 0.5$) times greater than the Monte Carlo method after $N = 2^{23}$ sample trajectories. In fact, the accuracy of the QMC particle simulation was between 40 and 2,400 times greater than even the absorption weighted Monte Carlo (AWMC) method, which had a lower variance than the traditional test particle method. The QMC particle simulation, however, had a greater computational cost per trajectory than the test particle Monte Carlo method. The reduction in simulation time for the QMC particle simulation, while still impressive, was not as substantial as the increase in accuracy. As an example, the QMC particle simulation reached a reference accuracy ($\epsilon \approx 10^{-4}$) between 6.2 ($L = 5$) and 1,800 ($L = 0.5$) times faster than the test particle Monte Carlo method.

In general, the increase in accuracy and speed achieved by the QMC particle simulation diminished as the duct became narrower (*i.e.* as L increases). To better quantify this performance loss, a power law model was fitted to the error convergence data of the QMC particle simulations with the model exponent serving as an empirical measure of convergence rate. The QMC particle simulation using the Halton and Niederreiter sequences consistently had the fastest convergence rate for all the duct

geometries tested, while the BCF-3 sequence was the slowest of the low-discrepancy sequences. The QMC particle simulation using the BCF-3 sequence, however, had the smallest power law constant, implying that it had the best initial accuracy of all the particle methods after a relatively small number of samples. For wider duct geometries with $L \leq 2.5$, the error convergence rate of the QMC particle simulations using the Halton and Niederreiter sequences was found to be $\mathcal{O}\left(N^{-0.92}\right)$, where N was the number of samples. As the duct length to height ratio L increased, the convergence rate of all the QMC particle simulations steadily declined until reaching a level between $\mathcal{O}\left(N^{-0.55}\right)$ and $\mathcal{O}\left(N^{-0.60}\right)$ at $L = 10$. The QMC particle simulation therefore converged at a faster rate than the Monte Carlo methods for all the geometries tested in this investigation.

The QMC particle simulation demonstrated a faster convergence rate than the Monte Carlo methods, albeit with a higher computational cost per sample. Consequently, there existed a critical error level E_{crit} at which the QMC particle simulation method became faster than the test particle Monte Carlo method. The QMC particle simulation method was then found to be a more efficient particle method for reaching any error level less than E_{crit} because of its superior error convergence rate. In general, as the desired simulation error decreases below E_{crit}, the time savings associated with QMC particle methods become even greater. The simulation of low-speed non-equilibrium gas flows, which are common in fluidic MEMS, often requires the bulk velocity field to be resolved to extremely small error levels relative to the random speed of the particles. It was thus important to measure E_{crit} in this investigation in order to understand the potential savings that the QMC method offers to the future design of a fluidic MEMS particle simulation.

The QMC particle simulation using the BCF-3 sequence was found to have the

largest critical error, for $L \leq 7.5$; and thus, was faster than the traditional test particle Monte Carlo method for the largest range of simulation accuracies. In fact, the critical error E_{crit} was greater than 18% for the QMC particle simulation using the BCF-3 sequence when $L \leq 6$. For $7.5 < L \leq 10$, the QMC particle simulation using the Niederreiter sequence in base 2 possessed the largest critical error E_{crit}, which remained between 0.1% and 1%. The QMC simulation, using either the BCF-3 or Niederreiter sequences, was therefore the preferable particle method for most simulation accuracies of practical interest.

The loss in performance observed in the QMC simulation when the duct became narrower was due to the increase in the number of particle moves required for each sample trajectory. Each particle move represented an independent dimension of the problem and, therefore, was generated by an independent dimension of the low-discrepancy sequence. It is well-documented throughout the QMC literature that the QMC method tends to converge slower in practice when the required dimension of the low-discrepancy sequence increases. To better understand this loss in performance due to the dimension of the low-discrepancy sequence, a measure of the non-physical correlation present in the QMC particle simulation was proposed in Section 6.4. This measure, denoted by N_{min}, indicated the length of the low-discrepancy sequence necessary for the two dimensions of the sequence to be considered as uncorrelated as a random sequence of equivalent length. Regardless of the type of low-discrepancy sequence, the value of N_{min} was shown to steadily increase with the sequence dimension. Ideally, one should try to collect more samples for the QMC method than N_{min} in order to maintain a physically consistent approximation. This was not shown to be always feasible for the QMC particle simulations of free molecular flow in the narrower duct geometries. In particular, N_{min} was found to be 10 times greater

than the $N = 2^{23}$ samples collected here for the QMC particle simulation of the $L = 2.5$ duct geometry, when a 55 dimensional Niederreiter sequence in base 2 was used. Because the QMC particle simulation demonstrated faster convergence for all duct geometries tested, including $L > 2.5$, the mere presence of correlation was not enough to condemn the method. With respect to its future use, the correlation analysis introduced in Section 6.4 is therefore best used to identify which simulation steps are most likely to produce physically inconsistent behavior.

To mitigate the performance loss observed in the QMC particle simulation of narrower duct geometries, a hybrid quasi-Monte Carlo/Monte Carlo (QMC/MC) method was developed in Section 6.5. The hybrid QMC/MC method reduced the necessary dimension of the low-discrepancy sequence by generating only a fraction of the particle moves with the sequence. The remaining particle moves were then generated using the traditional test particle Monte Carlo method. Since the initial particle moves had the greatest impact on the sample trajectory score, these were generated by the low-discrepancy sequence in the hybrid QMC/MC method. The hybrid QMC/MC method achieved nearly the same error convergence as the original QMC particle simulation while generating less than a third of the particle moves with the low-discrepancy sequence. As a consequence, the hybrid QMC/MC method was found to be two to five times faster. These results are only preliminary; however, they do suggest that the range of applicability for the QMC particle simulations can be further extended by considering additional dimension reduction techniques.

Several new[2] contributions to the improved simulation of microscale gas flow were developed by the author, during the course of this investigation. These contributions are highlighted in the following list:

[2]At least to the knowledge of the author.

- **Approach 1:** Empirical corrections to the Navier-Stokes solution
  - Two new empirical models were designed to correct the Navier-Stokes solution for Couette and Poiseuille flows in the transition regime ($0.01 \leq Kn \leq 10$). The new empirical models are similar to models developed by Karniadakis and Beskok [15, 69] and Bahukudumbi *et. al.* [11, 12], except that the database of known non-equilibrium solutions used in their construction was generated by the DSMC method instead of the linearized Boltzmann equation.
  - A new technique was implemented in the construction of these empirical models to achieve a more uniform accuracy throughout the transition regime. In particular, the sensitivity of the macroscopic flow quantities to changes in the empirical models was included in the Levenberg-Marquardt non-linear data-fitting of the known non-equilibrium solutions. The use of this sensitivity analysis represents an improvement in the model construction over earlier efforts by McNenly, Gallis and Boyd in [111, 112].
  - The testing and evaluation of these new empirical models was also unique to this investigation. Previous studies of such models [11, 12, 15, 69] have focused on the ability of the corrected Navier-Stokes solution to reproduce the known non-equilibrium solutions used in their construction. In contrast, the actual predictive capabilities of the empirical models to capture non-equilibrium flows outside the database of known solutions was evaluated in this investigation.
- **Approach 2:** Quasi-Monte Carlo (QMC) particle simulation
  - A new implementation of the low-discrepancy Weyl-Richtmyer sequence

(referred to as the BCF-3 sequence) was proposed in this investigation for use in the QMC particle simulation. For the simulation of free molecular duct flow, the BCF-3 sequence was shown to offer marked improvement (in terms of the single sample error and the critical error E_{crit}) when compared to other types of Weyl-Richtmyer sequences found in the literature.

– A QMC particle simulation was developed to approximate the conductance probability of free molecular flow in a two dimensional duct. While it does not appear as if this specific application has been studied previously,[3] the new QMC particle simulation is based on the method of Sarkar and Prasad [153] for a model transport problem. The QMC simulation developed in this investigation is also similar (in a mathematical sense) to QMC applications found in light ray-tracing [71, 72] and radiation transport [74]. Unlike these other applications from the literature, the QMC simulation of free molecular flow is more challenging because the lack of natural particle absorption increases the problem dimension.

– This new QMC particle simulation was also extensively tested on a large number of free molecular duct geometries, and was found to achieve a near-linear error convergence rate when the duct to length ratio was less than three. In contrast with many of the applications found in the QMC literature, which often only present a single case, the error convergence data was combined for all the test cases to illustrate the impact of the duct length on the performance of the QMC simulation. Further, this investigation was focused more on the practical advantages of the QMC

[3]There is no discussion of a QMC particle simulation for free molecular flow in the references cited in this thesis, nor is there mention in any of the proceedings of the International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing (1994-2006).

particle simulation over traditional Monte Carlo methods, in terms of the computational speedup available and the range of applicability.

– A new correlation measure was introduced in this investigation to quantify the extent of the non-physical correlation present between the dimensions of the low-discrepancy sequences used in the QMC particle simulations. While the correlation problems of the low-discrepancy sequences are well known [116, 146], there has not been any previous discussion about the sequence length required for the effects of non-physical correlation to be considered negligible. The new correlation measure serves to fill this void by calculating the minimum sequence length necessary for two dimensions of a low-discrepancy sequence to be considered as uncorrelated as a pseudo-random sequence of equivalent length.

– A new hybrid quasi-Monte Carlo/Monte Carlo (QMC/MC) particle simulation, based on the "mixed strategy" of Spanier [167], was proposed in this investigation to avoid the problems associated with the low-discrepancy sequences in a large number of dimensions. The hybrid QMC/MC particle simulation was then shown to achieve the same accuracy as the QMC particle simulation with a computational cost that was 2 to 4.5 times lower.

7.2 Future Work

The QMC particle simulation developed in Chapter V of this investigation is more computationally efficient than the traditional Monte Carlo methods under certain conditions of free molecular flow. Unfortunately, the QMC method was not yet shown to be a viable alternative to the DSMC method of Bird for general non-

equilibrium flows, in its current form described herein. The conclusions drawn from this investigation, however, do suggest several topics for future research that will extend the QMC particle simulation to a greater number of applications. These topics are briefly outlined in this section, and include: (i) improvements to the construction of the BCF-k sequence; (ii) the development of the QMC simulation for applications with natural particle absorption; and (iii) reductions in the dimension of the low-discrepancy sequences used in the QMC particle simulation.

7.2.1 Further improvements to the BCF-k sequences

Considering each low-discrepancy sequence tested in Section 6.3, the BCF-3 sequence had the largest value for the critical error E_{crit} for the QMC particle simulations when $L \leq 7.5$, as previously noted. The critical error, however, decreased rapidly when the duct became narrower (*i.e.* $L > 6$), which effectively limited the range of applications for the BCF-3 sequence. This reduction in the critical error E_{crit} was primarily due to the relatively slow convergence rate of $\mathcal{O}\left(N^{-0.55}\right)$ observed for the BCF-3 sequence when $L > 6$. By comparison, the QMC particle simulations using the Halton and Niederreiter ($b = 2$) sequences achieved faster error convergence rates between $\mathcal{O}\left(N^{-0.60}\right)$ to $\mathcal{O}\left(N^{-0.67}\right)$ for the same duct geometries. Because the Niederreiter sequence in base 2 had the lowest computational cost associated with its generation, the QMC particle simulation using the sequence had the largest value for the critical error E_{crit} when $L > 7.5$. The slower error convergence rate of the BCF-3 sequence is suspected to be caused by correlation problems between the sequence dimensions. While the Niederreiter sequence in base 2 was shown in Section 6.4 to have the most persistent correlation among the patterns considered, a general search of all possible dimension pairs revealed that the correlation in the

	Number of dimension pairs satisfying		
s	$N_{min} > 2^{21}$	$N_{min} > 2^{22}$	$N_{min} > 2^{23}$
50	3	3	2
100	34	22	16
150	115	80	63
200	219	162	124
250	317	231	172
300	586	448	351

Table 7.1: The number of dimension pairs of the BCF-3 sequence which have significant two dimensional correlation which persists over sequence lengths comparable to the maximum used by the QMC particle simulations.

BCF-3 sequence was actually much worse. In fact, existing within the first 50 dimensions of the BCF-3 sequence is a dimension pair that did not reach the same level of correlation as a random sequence until the sequence length (*i.e.* N_{min} (6.7)) became 25 billion. Further, the correlation persisted over 200 times longer than any of the other 50 dimensional low-discrepancy sequences tested in this investigation (see Figure 6.17).[4]

To demonstrate the extent of the correlation in an s-dimensional BCF-3 sequence, the dimension pairs that have a two dimensional correlation which persists beyond certain threshold lengths T (*i.e.* $N_{min} > T$) is shown in Table 7.1. Suppose a pair of irrational numbers in the set $\mathbf{z} = (z_1, \ldots, z_s)$ used to construct the BCF-3 sequence is known to yield significant two dimensional correlation. In such a case, one of the problematic irrational numbers in the correlated pair could simply be replaced without affecting the discrepancy of each one dimensional projection of the sequence.

[4]In Section 6.4, the largest values of N_{min} were presented for an exhaustive search of all dimension pairs of the Faure sequence. While not initially performed for the results in Figure 6.17, a subsequent search of all dimension pairs of the Halton sequence revealed that the largest values of N_{min} corresponded to all the twin prime bases presented in the figure. Similarly, the search of all dimension pairs of the Niederreiter sequence ($b = 2$) demonstrated that the largest values of N_{min} corresponded to the specific construction pattern tested in Figure 6.17, in most instances. For the few dimension pairs not appearing in the figure, the values of N_{min} remained bounded by the initial results; thus, they did not change the rate of growth of N_{min} as sequence dimension s increased.

This feature is unique only to the BCF-3 sequence. If this replacement strategy were adopted for the Halton or Niederreiter ($b = 2$) sequences, then the distribution of points in each one dimensional projection of the sequence would become increasingly less uniform as more construction elements were replaced. Therefore, there is practical motivation to further refine the heuristic process proposed in this investigation for determining the set of irrational numbers used to construct the BCF-3 sequence; that is, by allowing the removal of dimension pairs with persistent correlation.

A simple improvement would be to generate a large set $\mathbf{z}$ of irrational numbers for the BCF-3 sequence using Algorithm 4.1 and then calculate the extent of the two dimensional N_{min} for each pair of irrational numbers. The values of N_{min} greater than a prescribed threshold length T could then be easily searched, and an irrational number could be removed from the set $\mathbf{z}$ for each dimension pair exceeding the limit. The BCF-3 sequence constructed from the remaining irrational numbers in $\mathbf{z}$ would, as a result, have a better distribution of points in both the one and two dimensional projections of the sequence when the sequence length is greater than T. There is, however, no reason to limit future refinement of the BCF-k sequence to the bounding constant $k = 3$. Other small values of the bounding constant, such as $k = 4, 5, 6$, should also be considered as these BCF-k sequences may possess a much lower two dimensional correlation while only producing a modest increase in the one dimensional discrepancy bound. Furthermore, the elimination of dimensions which possess significant correlation between more than two dimensions may also yield a higher error convergence rate and should be explored as well.

As an initial foray into a more comprehensive search for an improved BCF-k sequence, a 300 dimensional BCF-5 sequence was constructed such that the absolute difference between any two distinct irrational numbers in the set $\mathbf{z}$ used to generate

the sequence was less than $7.8 \cdot 10^{-4}$. By preventing any two irrational numbers from becoming arbitrarily close, the correlation problems associated with the construction pattern for the BCF-k sequence in Figure 6.15(a) were avoided. Despite the removal of this single construction pattern being less thorough than the systematic removal of all dimension pairs with N_{min} greater than some threshold length T, the initial performance gains for the QMC particle simulation were nonetheless still promising. Specifically, this BCF-5 sequence achieved a faster error convergence rate than the BCF-3 sequence (*e.g.* between $\mathcal{O}\left(N^{-0.61}\right)$ to $\mathcal{O}\left(N^{-0.67}\right)$ when $L > 6$), which was comparable to the performance of the Halton and Niederreiter ($b = 2$) sequences. The critical error E_{crit} of this BCF-5 sequence also decreased much more slowly as the duct became narrower; however, it did not actually surpass the critical error of the BCF-3 sequence until $L \geq 10$. In light of these preliminary results, it seems likely that further improvements could be made for the BCF-k sequence by removing highly correlated sequence dimensions.

7.2.2 Free molecular flows with greater natural particle absorption

The performance of the QMC particle simulation suffered as the problem dimension increased, as previously noted. The probability that a particle reached an absorbing state during one move of the simulation directly affected the problem dimension. Stated more specifically, the higher the probability was of being absorbed during one move the smaller the problem dimension. In the simulation of free molecular duct flow, the absorbing states for particles inside the duct were simply the inlet and outlet. As the duct became narrower, the probability of a particle reaching the absorbing states (*i.e.* escaping through the inlet or outlet) decreased, which resulted in an increase in the problem dimension. The QMC particle simulation thus

converged at a slower rate as the duct length to height ratio L increases. If there was somehow a natural increase in the absorption probability which was independent of the duct geometry, then the performance losses associated with the QMC particle simulation would have diminished. There are, in fact, several applications of free molecular flow that possess a natural increase in particle absorption, and thus deserve further study.

One such application is the simulation of free molecular flow in a duct with partial wall accommodation. The QMC particle simulation developed in Chapter V assumed a fully diffuse wall; however, as the number of specular wall reflections increases, so too does the probability of a particle reaching the inlet or outlet. Note that if the all particle-wall collisions are specular, then the conductance probability of the duct is 100%. Another application, which has a similar increase in the escape probability of the particles, is the simulation of free molecular duct flow when the particles are accelerated by a body force in the direction of either the inlet or outlet. This type of flow condition occurs in low-density gas centrifuges and plasma propulsion systems. Given the natural increase in the escape probability, the QMC particle simulations of these two applications are then expected to yield a higher error convergence rate for a wider range of duct geometries.

In addition to the applications with an increase in the escape probability, there are some cases of free molecular flow in which the simulated particles are actually absorbed within the duct. An important engineering example occurs in the Low Pressure Chemical Vapor Deposition (LPCVD) process that is used in the manufacturing of computer chips and MEMS. For example, in the deposition of silicon dioxide (SiO_2) using silane (SiH_4) gas and oxygen, the probability that a silane molecule "sticks" to the silicon wafer after a collision is approximately 24% (at 400°C [184]). The presence

of this natural absorption of the simulated particles at the boundary walls reduces the effective dimension of the problem. In fact, the amount of natural absorption is sufficient to limit the effective dimension of the problem to less than 80 for the QMC particle simulation of the conductance probability of silane gas through a micro-scale silicon channel, regardless of length. The absorption of simulated particles need not be limited to just the wall boundaries. Any chemical reaction or ionization process in which the simulated particles interact with a background source (*i.e.* not other simulated particles) also increases the absorption probability. Therefore, the presence of these types of absorption processes in free molecular flow will also reduce the effective dimension of the problem.

7.2.3 Reducing the dimension of the low-discrepancy sequences

The hybrid QMC/MC method in Section 6.5 demonstrated some of the benefits of reducing the dimension of the low-discrepancy sequence required by the QMC portion of the simulation. Specifically, using the hybrid QMC/MC method, the same accuracy of the original QMC particle simulation was achieved in a fraction of the computation time. This reduction in computation time was entirely due to a lower cost associated with generating each sample trajectory, since the convergence rate of both the hybrid and original QMC methods was the same. While a lower cost per sample was of some benefit, an increase in the error convergence rate would be much more desirable because the performance gains would have a greater range of applicability. Given the substantial interest in dimension reduction techniques that actually achieve a higher error convergence rate, a few possible methods are briefly outlined here for future work.

A common way to reduce the necessary dimension of the low-discrepancy sequence

is simply to reuse each sequence dimension in a manner such that the independence of the simulation dimensions is properly maintained. Suppose one limited the dimension of the low-discrepancy sequence to s_{qmc} for a free molecular duct geometry requiring s particle moves (or independent dimensions of the problem), where $s = ks_{qmc}$ with the integer $k \geq 2$. There are three methods for reusing the s_{qmc} dimensions of the low-discrepancy sequence to generate the s particle moves in the simulation: (i) re-order the particles by position within the duct (based on the concept of Lécot [87], and Morkoff and Caflisch [115]); (ii) re-order the particles by a pseudo-random permutation (based on the concept of the "scrambled" hybrid QMC/MC method in [167]); and (iii) re-sample the distribution of particles within the duct. In each method, the sample trajectories are simultaneously calculated in sets of s_{qmc} particle moves. To avoid non-physical correlation between the sets of particle moves, the order in which each new particle move is generated from the low-discrepancy sequence is modified. There are, however, two drawbacks to these dimension reduction techniques that are not present in the original QMC formulation; namely, the total number of sample trajectories must be decided upon in advance, and the location of each trajectory must be stored in memory throughout the simulation.

Another dimension reduction technique involves the discretization of the free molecular duct geometry. For example, divide the computation domain of a narrow duct (*i.e.* the duct length to height ratio $L \gg 1$) into square cells with a side length equal to the duct height. The probability of a particle escaping a cell ($L_{cell} = 1$) during one move is nearly 40%, implying that a relatively small number of indepnendent particle moves are needed per cell. As noted in Chapter VI, the error convergence rate was between $\mathcal{O}\left(N^{-0.86}\right)$ and $\mathcal{O}\left(N^{-1.09}\right)$ for the QMC particle simulation of the $L = 1$ duct geometry. It is important that the transport of particles between cells

is simulated in such a manner as to ensure that no discontinuities are introduced into the sampling procedure. The presence of discontinuities in the QMC simulation typically implies the following (see Section 5.5): (i) the Koksma-Hlawka inequality cannot be used to establish the convergence and consistency of the method; and (ii) the convergence rate observed in practice is only slightly better than the Monte Carlo method. The normal particle transport through the computational cells of a DSMC simulation is discontinuous, unfortunately. The most critical design goal of a QMC particle simulation for a discretized computation domain is thus the elimination of these discontinuities. It may be possible to achieve this by adopting an alternative formulation of free molecular flow in which the distribution of particles fluxing across the cell boundaries is re-sampled after a fixed number of moves or time steps. The ability to discretize the computation domain is essential to any simulation of a complex geometry. If such a dimension reduction technique is shown to maintain a near-linear error convergence rate over a wide range of duct geometries, it would therefore represent a major milestone in development of a general QMC particle simulation to supplant the DSMC method.

APPENDICES

APPENDIX A

The van der Corput sequence (1935)

For the construction of the van der Corput sequence [178], along with the multi-dimensional low-discrepancy sequences presented in Appendices B-F, it is convenient to introduce the following two function definitions for $\vec{\xi}_b(n)$ and $\chi_b(n)$ and the accompanying notation. The first function $\vec{\xi}_b(n)$ represents the natural number n in base b as an infinite vector; that is $\vec{\xi}_b : \mathbb{N} \mapsto \mathbb{Z}_b^\infty$, where $\mathbb{Z}_b = \{0, 1, \ldots, b-1\}$ represents the set of possible digits in base b, and is referred to as the least residue system modulo b. Let $\vec{\xi}_b(n) = (\xi_{1,b}(n), \xi_{2,b}(n), \xi_{3,b}(n), \ldots)$ denote the vector components of the base b representation function, then $\vec{\xi}_b(n)$ is defined explicitly by

$$\xi_{m,b}(n) = \mathrm{mod}\left(\left\lfloor \frac{n}{b^{m-1}} \right\rfloor, b\right) \quad \text{for } m \geq 1 \text{ and } b \geq 2, \tag{A.1}$$

where $\lfloor \cdot \rfloor$ is the floor function, and is equal to the largest integer not greater than its argument. Note that the modulo function, as it is used in (A.1), is equivalent to the common computer programming definition. Specifically, for $x, b \in \mathbb{Z}$, with $b \geq 2$, the function is defined by

$$\mathrm{mod}(x, b) = y,$$

where $y \in \mathbb{Z}_b$ satisfies the congruence relationship $x \equiv y \pmod{b}$. As an example of its usage, the base b representation function number $\vec{\xi}_b(n)$ is found for the number

$n = 13$ in bases $b = 2, 3, 5$:

$$\begin{aligned} \vec{\xi}_2(13) &= (1, 0, 1, 1, 0, 0, \ldots) \\ \vec{\xi}_3(13) &= (1, 1, 1, 0, 0, \ldots) \\ \vec{\xi}_5(13) &= (3, 2, 0, 0, \ldots), \end{aligned}$$

where the ellipses denote an infinite sequence of zeros.[1]

The second function $\chi_b(n)$ is the radical inverse function in base b, which maps its argument from the natural numbers to the half-closed unit interval, stated more compactly $\chi_b : \mathbb{N} \mapsto [0, 1)$. For $b \geq 2$, the radical inverse function in base b is defined as

$$\chi_b(n) = \sum_{m=1}^{\infty} \xi_{m,b}(n) b^{-m}. \tag{A.2}$$

Note that $\xi_{m,b}(n) = 0$ for all $m > 1 + \log_b n$, thus the summation in (A.2) can be taken over a finite number of terms in practice. For the van der Corput sequence $S_C = (x_0, x_1, x_2, \ldots)$ in base b, the n^{th} term of the sequence is simply given by the inverse radical function such that $x_n = \chi_b(n)$. The first 16 points of the van der Corput sequence are calculated in Table A.1 in the bases $b = 2, 3, 5$.

Recall in Figure 3.8 that the star discrepancy of the van der Corput sequence S_C in base 2 appears to have a near linear convergence rate. In fact, for any base $b \geq 2$, the asymptotic convergence rate for the star discrepancy of the van der Corput sequences $D_N^*(S_C) = \mathcal{O}\left(N^{-1} \log N\right)$, as the number of points N tends toward infinity. While the specific base of the van der Corput sequence does not affect the asymptotic convergence rate, it does impact the implied constant in the Landau, or the big-$\mathcal{O}$, notation. The general behavior is such that the implied constant tends to increase,

[1] From the viewpoint of a programmer, it is impractical to treat the vector $\vec{\xi}_b(n)$ as infinite, especially when almost all the terms are zero. However, it is easier to develop general sequence formulas with the infinite vector representation because it avoids the need for conditional specifications on the size of $\vec{\xi}_b(n)$ as n and b change.

n	$\chi_2(n)$	$\chi_3(n)$	$\chi_5(n)$
0	0.0000	0.0000	0.0000
1	0.5000	0.3333	0.2000
2	0.2500	0.6667	0.4000
3	0.7500	0.1111	0.6000
4	0.1250	0.4444	0.8000
5	0.6250	0.7778	0.0400
6	0.3750	0.2222	0.2400
7	0.8750	0.5556	0.4400
8	0.0625	0.8889	0.6400
9	0.5625	0.0370	0.8400
10	0.3125	0.3704	0.0800
11	0.8125	0.7037	0.2800
12	0.1875	0.1481	0.4800
13	0.6875	0.4815	0.6800
14	0.4375	0.8148	0.8800
15	0.9375	0.2593	0.1200

Table A.1: The first 16 points constructed for the van der Corput sequences in bases $b = 2$, 3 and 5.

as the base of the van der Corput sequence increases. Stated more precisely, Faure [45] establishes the following relationship between the base b and the asymptotic behavior of the van der Corput sequence:

$$\overline{\lim_{N\to\infty}} \frac{N D_N^* (S_C)}{\log N} = \begin{cases} \dfrac{b^2}{4(b+1)\log b} & \text{for even } b, \\ \dfrac{b-1}{4\log b} & \text{for odd } b. \end{cases} \tag{A.3}$$

Note that $\overline{\lim}_{N\to\infty} S(n)$ represents the upper limit (or limit superior) of a sequence $S(n)$, where n denotes the element number of the sequence. The mathematical definition of the upper limit is stated as follows. For some real constant k, the upper limit of the sequence $S(n)$

$$\overline{\lim_{N\to\infty}} S(N) = k$$

exists if, for every $\epsilon > 0$, $|S(n) - k| < \epsilon$ for infinitely many values of n and if no number larger than k has this property.

When using quasi-Monte Carlo integration, it is desirable to choose a low-discrepancy sequence with the smallest asymptotic constant for the star discrepancy. This

ensures that the error bound given by the Koksma-Hlawka inequality (3.5) is as small as possible. From the result (A.3) of Faure, the van der Corput sequence in base 3 achieves the smallest possible bounding constant. It is interesting to note that one can construct a one dimensional low-discrepancy sequence with an asymptotic bound that is even smaller than the result in (A.3) with a slight modification of the van der Corput sequence. Let σ denote a specific permutation of the possible digits in base b; that is, $\sigma : \mathbb{Z}_b \mapsto \mathbb{Z}_b$ is a one-to-one mapping of the digits $\{0, 1, \ldots, b-1\}$ onto themselves. The generalized van der Corput sequence in base b using the permutation σ is then defined by

$$x_n = \sum_{m=1}^{\infty} \sigma\big(\xi_{m,b}(n)\big) b^{-m},$$

where x_n is the n^{th} element of the sequence. The lowest proven star discrepancy bound on a generalized van der Corput sequence (at least at the time of publication of [127]) belongs to a special permutation of the base 12 sequence constructed by Faure in [45], which yields an upper limit of

$$\overline{\lim_{N\to\infty}} \frac{ND_N^*}{\log N} \frac{1919}{3454 \log 12}. \tag{A.4}$$

The upper limit (A.4) is over three times smaller than the minimum for the base $b = 3$ sequence in (A.3).

APPENDIX B

The Weyl-Richtmyer Sequence (1916/1951)

First proposed in [186], Weyl established that this sequence yields a mathematically consistent integral approximation when used in the QMC method. Richtmyer further developed the sequence as part of the first published QMC simulation[1] in [148]. Even more importantly, Richtmyer also established that the QMC approximation using this sequence converges to the true solution almost linearly. In recognition of both contributions, this sequence is referred to as the "Weyl-Richtmyer sequence" throughout this investigation.

The Weyl-Richtmyer sequence is remarkably simple in its construction (see [127]). In one dimension, the n^{th} term of the sequence is defined as

$$x_n = [nz], \tag{B.1}$$

where z is an irrational number. The notation $[nz]$ is sometimes read as "nz modulo 1," and represents the fractional part of its argument. An alternative description is $[nz] = nz - \lfloor nz \rfloor$, where $\lfloor \cdot \rfloor$ is the floor function, which yields the greatest integer not larger than the argument. Thus, the operation $[\cdot] : \mathbb{R} \mapsto [0, 1)[x]_{\mathrm{mod}\,1}$.

The extension of the Weyl-Richtmyer sequence to more dimensions naturally

[1] In fact, it was Richtmyer [148] who first coined the term "quasi-Monte Carlo integration."

follows from the one dimensional sequence, and now the n^{th} vector is defined as

$$\mathbf{x}_n = [n\mathbf{z}], \tag{B.2}$$

where the application of $[\cdot]$ to an s-dimensional vector, maps $\mathbb{R}^s \mapsto [0,1)^s$ in a normal way,

$$[n\mathbf{z} = ([nz_1], [nz_2], \ldots, [nz_s]) .$$

The vector $\mathbf{z}$ consists of irrational numbers that are linearly independent over the rationals. This constraint implies that $\mathbf{z}$ must be chosen such that it is impossible to find a solution in the integers to the following linear equation

$$c_1 z_1 + c_2 z_2 + \cdots + c_s z_s = c_{s+1}, \quad c_i \in \mathbb{Z}.$$

Further discussion of the necessary conditions for certain sets of irrational numbers to be considered linearly independent over the rationals is given in Section 4.2 and in Besicovitch's Theorem [14].

Weyl [186] established that the infinite sequence generated by (B.2) is both *uniformly distributed modulo 1* (u.d. mod 1) and *well-distributed modulo 1* (w.d. mod 1). A review of Weyl's results[2] is provided by Kuipers and Niederreiter in Chapter 1 of [85]. Informally, an infinite sequence is u.d. mod 1 if the fraction of points in every subinterval in $[0,1)^s$ equals the volume of the subinterval. This is equivalent to saying the extreme discrepancy (3.22) of the infinite sequence tends to zero. An infinite sequence is w.d. mod 1 if after removing the first k elements of the sequence; the resulting subsequence is u.d. mod 1, for $k = 0, 1, 2, \ldots$. Therefore, an infinite sequence that is w.d. mod 1 is also u.d. mod 1 with an additional uniformity constraint on the order of the sequence elements.

[2]Kuipers and Niederreiter [85] provide a very thorough description (and useful translation) of Weyl's original work [186] which is in German.

In addition to creating the definitions of u.d. mod 1 and w.d. mod 1, which were critical to the development of the concept of discrepancy, Weyl proved two theorems that are of the utmost importance to quasi-Monte Carlo integration. These theorems establish that the QMC method of integration is a mathematically consistent approximation, and are stated as follows in [85].

Theorem B.1 *The sequence* $\{\mathbf{x}_n\}$, $n = 1, 2, \ldots$ *is u.d. mod 1 if and only if for every real-valued continuous function* f *defined on the closed interval* $[0,1]^s$ *we have*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(\mathbf{x}_n) = \int_{[0,1]^s} f(\mathbf{u})\, d\mathbf{u}. \tag{B.3}$$

Theorem B.2 *The sequence* $\{\mathbf{x}_n\}$, $n = 1, 2, \ldots$ *is u.d. mod 1 if and only if for every real-valued continuous function* f *defined on the closed interval* $[0,1]^s$ *we have*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=k+1}^{k+N} f(\mathbf{x}_n) = \int_{[0,1]^s} f(\mathbf{u})\, d\mathbf{u} \quad \text{uniformly in } k = 0, 1, 2, \ldots. \tag{B.4}$$

Theorems B.1 and B.2 prove that sampling a continuous function f with the Weyl-Richtmyer sequence (B.2) produces a consistent approximation of the integral of f.[3] Moreover, the theorems of Weyl establish the necessary and sufficient conditions for any low-discrepancy sequence used in QMC method to yield a consistent numerical approximation. The ideas of Weyl served as the foundation on which the discrepancy measure of a sequence is developed and ultimately led to the Koksma-Hlawka inequality, which is the cornerstone of the QMC method.

The Koksma-Hlawka inequality (3.5) establishes the upper error bound on error of QMC integration. For a function with bounded variation, the rate at which the star discrepancy of a sequence tends to zero determines the rate of error convergence of the integration method. The convergence rate of the extreme discrepancy of a one

[3]The concept of consistency is an essential property of any accurate numerical method. For a more detailed discussion of consistency as it relates to the simulation of fluid flows refer to [63].

dimensional Weyl-Richtmyer sequence $S(z)$ is given by (see Corollary 3.5 in [127])

$$D_N\left(S\left(z\right)\right) < G(k)N^{-1}\log(N+1) \quad \text{for all } N \geq 1, \tag{B.5}$$

where z is the irrational constant in (B.1), $G(k) = 2/\log 2$ for $k = 1, 2, 3$ and $G(k) = (k+1)/\log(k+1)$ for $k \geq 4$. In (B.5), the constant k is the smallest integer greater than or equal to the coefficients in the continued fraction expansion of z. The basic properties of continued fractions are covered in [78, 131], and a more thorough review of the mathematical theory is available in [75].

For an s-dimensional Weyl-Richtmyer sequence $S(\mathbf{z})$, where $\mathbf{z}$ is the vector generating the sequence in (B.2), there is a probabilistic bound on the extreme discrepancy. The probabilistic bound due to Schmidt [155] states that for every $\epsilon > 0$, $D_N(S(\mathbf{z})) = O\left(N^{-1}(1+\log N)^{s+1+\epsilon}\right)$ for almost all $\mathbf{z} \in \mathbb{R}^s$, where the exceptions are contained in a set with Lebesgue measure zero. If the vector $\mathbf{z}$ consists of algebraic numbers that are linearly independent over the rationals, then a deterministic bound $D_N(S(\mathbf{z})) = O\left(N^{-1+\epsilon}\right)$ for every $\epsilon > 0$ is given by Niederreiter [123]. Note that the extreme discrepancy is an upper bound to the star discrepancy. Thus, the results for the extreme discrepancy also apply to the Koksma-Hlawka inequality (3.5) and demonstrate that the error convergence of QMC integration with the Weyl-Richtmyer sequence is nearly linear.

There is no restriction on the choice of irrational numbers for the vector $\mathbf{z}$ in (B.2) other than their linear independence over the rationals. The low-discrepancy sequence need not be constructed solely from quadratic numbers (as in Chapter IV and [68, 148]); other families of irrational numbers can be considered. For $m, n \in \mathbb{Z}$ and $r \in \mathbb{Q}$, these families of irrational numbers include the following: (i) $n^{1/m}$ where n is not the m^{th} power of an integer; (ii) $\log_n m$ where $\gcd(m, n) \neq m$ or n; (iii) e^r for

n	$x_1(n)$	$x_2(n)$	$x_3(n)$
0	0.0000	0.0000	0.0000
1	0.4142	0.7321	0.2361
2	0.8284	0.4641	0.4721
3	0.6569	0.9282	0.9443
4	0.3137	0.8564	0.8885
5	0.6274	0.7128	0.7771
6	0.2548	0.4256	0.5542
7	0.5097	0.8513	0.1084
8	0.0193	0.7025	0.2167
9	0.0387	0.4050	0.4334
10	0.0773	0.8100	0.8668
11	0.1547	0.6200	0.7336
12	0.3094	0.2401	0.4672
13	0.6188	0.4801	0.9344
14	0.2375	0.9602	0.8689
15	0.4750	0.9204	0.7377

Table B.1: The first 16 points constructed for a three dimensional Weyl sequence using the fractional parts of the irrational numbers $\sqrt{2}$, $\sqrt{3}$ and $\sqrt{5}$.

distinct $r \neq 0$; (iv) $\cos r$ for distinct $r \neq 0$; and (v) $\tan r$ for distinct $r \neq 0$. However, the use of algebraic irrational numbers (*i.e.* those that are roots of a polynomial with integer coefficients), allows for the deterministic error bound on the extreme discrepancy of Niederreiter to be applied. For further reference, the first 16 points of a three dimensional Weyl-Richtmyer sequence generated from $\mathbf{z} = \left(\sqrt{2}, \sqrt{3}, \sqrt{5}\right)$ are given in Table B.1.

APPENDIX C

The Halton Sequence (1960)

This multi-dimensional low-discrepancy sequence was first investigated by Halton in [56]. The s-dimensional Halton sequence is actually constructed from s distinct one dimensional van der Corput sequences (see Appendix A) that satisfy a certain independence condition. Specifically, the n^{th} element of the Halton sequence $S_H = \mathbf{x}_0, \mathbf{x}_1, \ldots \in \overline{I}^s$ is defined by

$$\mathbf{x}_n = (\chi_{p_1}(n), \chi_{p_2}(n), \ldots, \chi_{p_s}(n)),$$

where $p_1, \ldots, p_s$ are pair-wise relatively prime integers,[1] and $\chi_b(n)$ is the inverse radical function defined in (A.2) for the van der Corput sequence in base b. Each dimension of the Halton sequence is therefore an independently generated van der Corput sequence.

There exists an explicit upper bound on the star-discrepancy of the s-dimensional Halton sequence S_H generated using the pair-wise relatively prime set of integers $p_1, \ldots, p_s$ for the bases of the van der Corput sequence. From Theorem 3.6 in [127],

$$D_N^*(S_H) < \frac{s}{N} + \frac{1}{N} \prod_{i=1}^{s} \left(\frac{p_i - 1}{2 \log p_i} \log N + \frac{p_i + 1}{2} \right) \quad \text{for all } N \geq 1. \tag{C.1}$$

[1] A set of integers $p_1, \ldots, p_s$ is said to be *pair-wise relatively prime* if $\gcd(p_i, p_j) = 1$ for all $1 \leq i, j \leq s$ except $i = j$.

Adopting the same form used for the other low-discrepancy sequences, the result in (C.1) is re-written to yield

$$D_N^* \leq C_s^H \frac{(\log N)^s}{N} + \mathcal{O}\left(N^{-1}(\log N)^{s-1}\right), \tag{C.2}$$

where the bounding constant C_s^H is given by

$$C_s^H = \prod_{i=1}^{s} \frac{p_i - 1}{2 \log p_i}. \tag{C.3}$$

It is clear that the bounding constant C_s^H (C.3) achieves the smallest value when the pair-wise relatively prime bases are chosen to be as small as possible. With respect to the theoretical bound on the discrepancy of the Halton sequence, it is therefore optimum to select the bases $p_1, \ldots, p_s$ to be the smallest s prime numbers, which is the standard construction for this sequence. As an example, the standard construction[2] of the three dimensional Halton sequence is given in Table C.1.

[2]A word of caution is in order for the actual calculation of the Halton sequence because the algorithms outlined in [56, 57] may produce numerically unstable results. Techniques to avoid these stability problems are presented in Section 4.3 and by Fox in [48].

n	$x_{1,n}$	$x_{2,n}$	$x_{3,n}$
0	0.0000	0.0000	0.0000
1	0.5000	0.3333	0.2000
2	0.2500	0.6667	0.4000
3	0.7500	0.1111	0.6000
4	0.1250	0.4444	0.8000
5	0.6250	0.7778	0.0400
6	0.3750	0.2222	0.2400
7	0.8750	0.5556	0.4400
8	0.0625	0.8889	0.6400
9	0.5625	0.0370	0.8400
10	0.3125	0.3704	0.0800
11	0.8125	0.7037	0.2800
12	0.1875	0.1481	0.4800
13	0.6875	0.4815	0.6800
14	0.4375	0.8148	0.8800
15	0.9375	0.2593	0.1200

Table C.1: The first 16 points constructed for a three dimensional Halton sequence with prime bases $p_1 = 2$, $p_2 = 3$, and $p_3 = 5$.

APPENDIX D

The Sobol' Sequence (1967)

The low-discrepancy sequence of Sobol' [161] uses primitive polynomials in $\mathbb{F}_2[x]$ to construct binary bit-masks, which uniquely permute the van der Corput sequence in base 2 for each dimension of the Sobol' sequence. Refer to [96] for a thorough discussion of primitive polynomials over finite fields and how they are calculated. Each dimension of the sequence requires a unique primitive polynomial in $\mathbb{F}_2[x]$. Using the primitive polynomial with the lowest possible degree, Sobol' [161] establishes the following bound on the star discrepancy

$$D_N^* \leq C_s^S \frac{(\log N)^s}{N} + O(\frac{(\log N)^{s-1}}{N}), \tag{D.1}$$

where

$$C_s^S = \frac{2^\alpha}{s!(\log 2)^s}. \tag{D.2}$$

Here, the exponent α is a function of s and is bounded by

$$K \frac{s \log s}{\log \log s} \leq \alpha \leq \frac{s \log s}{\log 2} + O(s \log \log s),$$

where $K > 0$. The implied coefficient C_s^S in the Sobol' sequence bound (D.2) does grow super-exponentially as $s \to \infty$; however, the growth is still not as fast as the bounding constant C_s^H (C.3) for the Halton sequence. Another advantage of the

Sobol' sequence stems from the fact that the entire process of generating the sequence elements consists of logical binary operations which can be performed very efficiently on a modern computer. In fact, by ordering the necessary operations appropriately, great computational savings can be achieved when 32 logical binary operations are concatenated into a single bit-wise operation on a 32-bit computer word.

Each dimension of the sequence is generated from a unique primitive polynomial in $\mathbb{F}_2[x]$, which, in turn, defines a set of bit-masks used to build the sequence. The n^{th} member of the van der Corput sequence in base 2 is a reflection of the binary representation of n around the decimal point (*i.e.* $\vec{\xi}_2(n)$ using the definition in (A.1)). These bit-masks are set by the binary representation of n and serve to shuffle the original van der Corput sequence in base 2. The bit-masks are generated from the linear recurring sequence associated with each primitive polynomial [96]. Because these polynomials are primitive in $\mathbb{F}_2[x]$, the period length of linear recurring sequences used to generate the bit-masks is the maximum possible. The construction of the Sobol' sequence remains rather opaque compared to the more traditional mathematics found in other low-discrepancy sequences. Because the Sobol' sequence is nearly the same as the special construction of Niederreiter for a (t, s)-sequence in base 2 (see [125, 126, 127] and Appendix F), it is perhaps easier to consider its construction using the theory provided for the (t, s)-sequences. However, the sample construction of the Sobol' sequence presented here follows the original approach of Sobol' [161] and the more recent algorithms developed in [18, 147], rather than adopting the framework of the (t, s)-sequences.

The construction of a multi-dimensional Sobol' sequence is independent in each dimension; for convenience, only one dimension is initially considered here to prevent the need for an additional subscript or superscript in the notation. Furthermore, the

construction is limited to the first 2^k sequence members to avoid the need for infinite sums, where k is any positive integer. The first step is to select a primitive polynomial $p(x) \in \mathbb{F}_2[x]$ of degree d defined by

$$p(x) = x^d + a_1 x^{d-1} + \cdots + a_{d-1}x + 1, \tag{D.3}$$

where the coefficients $a_i \in \{0, 1\} = \mathbb{F}_2$ for $1 \leq i \leq d-1$. A table of primitive polynomials in $\mathbb{F}_2[x]$ is given in [96] (see pp. 384–398). Similar to the selection of the smallest possible primes for the Halton sequence, the best bounding constant C_s^S (D.2) for the Sobol' sequence occurs when the primitive polynomials have the smallest degree possible. The number of primitive polynomials in $\mathbb{F}_2$ of degree d is equal to $\phi\left(2^d - 1\right)/d$, where ϕ is Euler's totient function [131].

The second step is to generate the necessary bit-masks to produce the sequence. In the English translation of Sobol' [161], these bit-masks used to construct the sequence are referred to as "direction numbers." The direction numbers are rational fractions between zero and one that are represented exactly by a finite number of bits. That is to say, when the fraction a/b is represented in its reduced form (*i.e.* $\gcd(a, b) = 1$), the denominator must be a power of two. The coefficients $(a_1, \ldots, a_d)$ from the primitive root $p(x)$ in (D.3) form a d-term linear recurrence for calculating the direction numbers v_i given by

$$v_i = a_1 v_{i-1} \oplus a_2 v_{i-2} \oplus \cdots \oplus a_{d-1} v_{i-d+1} \oplus v_{i-d} \oplus \left(v_{i-d}/2^d\right) \quad \text{for } i > d, \tag{D.4}$$

where $\oplus$ denotes the bit-wise XOR operation performed relative to a fixed decimal point. In order to use the linear recurrence (D.4), one needs the initial direction numbers v_i for $1 \leq i \leq d$. These are defined by $v_i = m_i/2^i$ for $1 \leq i \leq d$, with the necessary conditions that m_i is odd and $0 < m_i < 2^i$. Ideally, one should select distinct m_i for each primitive root used to generate a dimension of the sequence.

This precaution diminishes the amount of correlation between the dimensions for the initial elements of the Sobol' sequence. Sobol' [162] also proposes further restrictions on the selection of m_i that ensure additional uniformity properties which are beyond the scope of this investigation.

Finally, the third step in the construction of the Sobol' sequence is to generate the actual sequence members from the direction numbers v_i, for $1 \leq i \leq k$. Using a finite analogue of the base b representation vector defined in (A.1) for the van der Corput sequence, let $\vec{\xi}_2(n) = (\xi_{1,2}(n), \ldots, \xi_{s,2}(n))$ denote the base 2 representation of an integer n. Let $n = b_k \ldots b_2 b_1$ denote the binary representation of the integer n. Then the n^{th} element of the Sobol' sequence x_n is given by

$$x_n = \xi_{1,2}(n) \cdot v_1 \oplus \xi_{2,2}(n) \cdot v_2 \oplus \cdots \oplus \xi_{k,2}(n) \cdot v_k. \tag{D.5}$$

Recall that this construction only applies to the first 2^k elements of the Sobol' sequence because the binary representation $\vec{\xi}_2(n)$ contains only the k least significant bits of n. The direction numbers v_i do not change throughout the sequence construction, and therefore need only be generated once. Antonov and Saleev [2] propose using the binary Gray code representation of the integer n to reduce the calculation in (D.5) to a single XOR operation. The significant computational savings and implementation of the Gray code modification is discussed in Section 4.3 for the algorithmic implementation of the Niederreiter sequence in base 2. In order to produce a multi-dimensional Sobol' sequence, the aforementioned procedure is simply repeated for each dimension using a distinct primitive polynomial in $\mathbb{F}_2$.

As an example, the direction numbers and a sequence member are calculated for the primitive polynomial $p(x) = x^2 + x + 1$ for the first 2^4 elements of the sequence (*i.e.* $k = 4$ i (D.4)). The linear recurrence for this specific polynomial is then given

by

$$v_i = v_{i-1} \oplus v_{i-2} \oplus v_{i-2}/4 \quad \text{for } i > 2. \tag{D.6}$$

To start the recurrence, initial values for $v_1 = m_1/2$ and $v_2 = m_2/4$ must be selected under the conditions that m_i is odd and $0 < m_i < 2^i$. With these conditions, the only valid selection for m_1 is the value 1, while m_2 is restricted to the values 1 and 3. For the remainder of this example, let $m_2 = 1$, thus $v_1 = 0.1000$ and $v_2 = 0.0100$ in binary representation of the direction numbers. Applying the linear recurrence in (D.6) yields the following direction numbers v_3 and v_4 (represented as binary fractions):

$$\begin{aligned} v_3 &= 0.0100 \oplus 0.1000 \oplus 0.0010 = 0.1110 \\ v_4 &= 0.1110 \oplus 0.0100 \oplus 0.0001 = 0.1011. \end{aligned}$$

To complete the illustration of the construction process, the 13^{th} element of the sequence x_{13} is calculated from the direction numbers $v_1, \ldots, v_4$ and the binary representation $\vec{\xi}_2(13) = (1, 0, 1, 1, 0, 0, \ldots)$ using (D.5). That is,

$$x_{13} = 1 \cdot 0.1000 \oplus 0 \cdot 0.0100 \oplus 1 \cdot 0.1110 \oplus 1 \cdot 0.1011 = 0.1101 = \tfrac{13}{16}.$$

The example is continued in Table D.1 for a three dimensional Sobol' sequence using the following primitive polynomials $p_i(x) \in \mathbb{F}_2[x]$ to generate the i^{th} dimension of the sequence:

$$\begin{aligned} p_1(x) &= x + 1 \\ p_2(x) &= x^2 + x + 1 \\ p_3(x) &= x^3 + x + 1 \end{aligned}$$

The initial direction numbers used for each primitive root are defined by: $m_1 = 1$ for dimension 1; (ii) $m_1 = 1$ and $m_2 = 1$ for dimension 2; and (iii) $m_1 = 1$, $m_2 = 3$ and $m_3 = 3$ for dimension 3.

n	$x_1(n)$	$x_2(n)$	$x_3(n)$
0	0.0000	0.0000	0.0000
1	0.5000	0.5000	0.5000
2	0.7500	0.2500	0.7500
3	0.2500	0.7500	0.2500
4	0.6250	0.8750	0.3750
5	0.1250	0.3750	0.8750
6	0.3750	0.6250	0.6250
7	0.8750	0.1250	0.1250
8	0.9375	0.6875	0.3125
9	0.4375	0.1875	0.8125
10	0.1875	0.9375	0.5625
11	0.6875	0.4375	0.0625
12	0.3125	0.3125	0.1875
13	0.8125	0.8125	0.6875
14	0.5625	0.0625	0.9375
15	0.0625	0.5625	0.4375

Table D.1: The first 16 points constructed for a three dimensional Sobol' sequence using primitive polynomials over $\mathbb{F}_2$: $p_1(x) = x + 1$, $p_2(x) = x^2 + x + 1$ and $p_3(x) = x^3 + x + 1$.

APPENDIX E

The Faure Sequence (1982)

The Faure sequence [46] is constructed from a single, one dimensional van der Corput sequence that is uniquely permuted for each sequence dimension to yield a multi-dimensional low-discrepancy sequence. Faure demonstrates this sequence, denoted by $S_F = (\mathbf{x}_0, \mathbf{x}_1, \ldots) \in \overline{I}^s$, achieves the following bound on its star discrepancy:

$$D_N^*(S_F) \leq C_s^F \frac{(\log N)^s}{N} + \mathcal{O}\left(N^{-1}(\log N)^{s-1}\right) \quad \text{with } C_s^F = \mathcal{O}(1). \tag{E.1}$$

The discrepancy bound (E.1) is an improvement, in an asymptotic sense, over the sequences of Halton (see Appendix C) and Sobol' (see Appendices C and D respectively). Unlike the Halton (C.3) and Sobol' (D.2) sequences, the bounding constant C_s^F is $\mathcal{O}(1)$ and therefore remains bounded in the limit as the sequence dimension s tends to infinity.

The construction of the s-dimensional Faure sequence requires a choice of a prime base q that is greater than or equal to s. To simplify the construction, the maximum sequence length is taken to be less than q^k elements, for some positive integer k. The assumption of a finite sequence length has no real impact on the final implementation, since it is impossible for a computer simulation to produce an infinite number of sequence elements. In addition, a matrix-vector notation is also adopted for the following discussion of the Faure sequence construction process. Let the vector $\mathbf{x}_n =$

$(x_{1,n}, \ldots, x_{s,n}) \in \overline{I}^s$ denote the n^{th} element of the Faure sequence in base q. Using a finite analogue of the base b representation vector defined in (A.1) for the van der Corput sequence, let $\vec{\xi}_q(n) = (\xi_{1,q}(n), \ldots, \xi_{s,q}(n))$ denote the base q representation of an integer n, for $0 \leq n \leq q^k - 1$. Next define the matrix $C = [c_{ij}] \in \mathbb{F}_q^{k \times k}$ as an upper-triangle matrix where the j^{th} column corresponds to the j^{th} row of Pascal's triangle modulo q; that is

$$c_{ij} = \begin{cases} \mod\left(\frac{(j-1)!}{(j-i)!(i-1)!}, q\right) & \text{for } 1 \leq i \leq j \leq k \\ 0 & \text{for } 1 \leq j < i \leq k. \end{cases} \tag{E.2}$$

Note that the non-zero elements in (E.2) can be determined equivalently by the binomial coefficient modulo q (*e.g.* $c_{ij} = \mod\left(\binom{j-1}{i-1}, q\right)$ for $j \geq i$); as a consequence, the matrix C is sometimes referred to as the binomial matrix. Finally, let the set of vectors $\mathbf{y}_n^{(m)} = (y_{1,n}^{(m)}, \ldots, y_{k,n}^{(m)}) \in \mathbb{F}_q^k$ for $1 \leq m \leq s$ denote the result from the matrix-vector multiplication given by

$$\mathbf{y}_n^{(m)} = C^{m-1} \vec{\xi}_q(n), \tag{E.3}$$

where $C^0 = I$ is the $k \times k$ identity matrix. It is important to note that the addition and multiplication operation in (E.3) are performed over the prime field $\mathbb{F}_q$. Since q is prime, the operations are equivalent to their standard definitions, except with the final result taken modulo q.

With the preceding matrix and vector definitions, each coordinate of the vector $\mathbf{x}_n$ representing the n^{th} element of the Faure sequence in base q is then determined by

$$x_{m,n} = \sum_{i=1}^{k} y_{i,n}^{(m)} q^{-i} \quad \text{for } 1 \leq m \leq s. \tag{E.4}$$

Note that the calculation in (E.4) is performed over the real numbers (*i.e.* standard arithmetic). With regards to the actual generation of the Faure sequence, it is

possible to calculate and store all $s-1$ powers of the binomial matrix C needed in (E.3). However, it is much more efficient in practice to store the single C matrix (E.2) and generate the set of vectors $\mathbf{y}_n^{(m)}$ for each sequence element $\mathbf{x}_n$ using the following iterative formulae [48]:

$$\begin{aligned} \mathbf{y}_n^{(1)} &= \vec{\xi}_q(n) \\ \mathbf{y}_n^{(m)} &= C\mathbf{y}_n^{(m-1)} \quad \text{for } 2 \leq m \leq s. \end{aligned} \tag{E.5}$$

Based on the first step in the iterative formulae in (E.5), the first coordinate $x_{1,n}$ of the Faure sequence in base q is the same as the van der Corput sequence in base q.

As an example, consider a Faure sequence in three dimensions, with a prime base $q = 3$ and a maximum number of digits $k = 4$. The aforementioned construction process can then be used to generate the first 3^4 members of the sequence. If a longer sequence is needed, a larger value for k must be selected. In this example, the matrix C (E.2) becomes

$$C = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

To generate the 13^{th} element $\mathbf{x}_{13}$ of the Faure sequence in base 3, one must first determine the base 3 representation of 13; that is,

$$\vec{\xi}_3(13) = (1,1,1,0)^T.$$

Then one repeatedly applies the iterative matrix-vector multiplication in (E.5) to

n	$x_1(n)$	$x_2(n)$	$x_3(n)$
0	0.0000	0.0000	0.0000
1	0.3333	0.3333	0.3333
2	0.6667	0.6667	0.6667
3	0.1111	0.4444	0.7778
4	0.4444	0.7778	0.1111
5	0.7778	0.1111	0.4444
6	0.2222	0.8889	0.5556
7	0.5556	0.2222	0.8889
8	0.8889	0.5556	0.2222
9	0.0370	0.5926	0.4815
10	0.3704	0.9259	0.8148
11	0.7037	0.2593	0.1481
12	0.1481	0.7037	0.2593
13	0.4815	0.0370	0.5926
14	0.8148	0.3704	0.9259
15	0.2593	0.1481	0.7037

Table E.1: The first 16 points constructed for a three dimensional Faure sequence with a prime base $q = 3$.

calculate the set of vectors $(\mathbf{y}_{13}^{(1)}, \mathbf{y}_{13}^{(2)}, \mathbf{y}_{13}^{(3)})$, which yields

$$\begin{aligned} \mathbf{y}_{13}^{(1)} &= (1,1,1,0)^T \\ \mathbf{y}_{13}^{(2)} &= (0,0,1,0)^T \\ \mathbf{y}_{13}^{(3)} &= (1,2,1,0)^T. \end{aligned}$$

Finally, this set of vectors is used in (E.4) to generate each coordinate of the 13^{th} element of the Faure sequence; specifically,

$$\mathbf{x}_{13} = (\tfrac{13}{27}, \tfrac{1}{27}, \tfrac{16}{27}).$$

For further reference, the calculation of the first 16 points of the Faure sequence in base 3 is provided in Table E.1.

APPENDIX F

The Niederreiter (t, s)-Sequence (1987)

Niederreiter, in [125], introduced the concept of a (t, s)-sequence in base q in order to establish a systematic theory connecting the construction of a low-discrepancy sequence with its discrepancy bound. The integer $t \geq 0$ can be viewed informally as a measure of the strength of the uniformity condition of the sequence, where a smaller value of t indicates that the (t, s)-sequence satisfies a stronger uniformity condition; and the integer s is the dimension of the sequence. Strictly speaking, a (t, s)-sequence does not define a specific sequence; rather, it is a classification system based on the general construction properties of a sequence. For example, the van der Corput, Sobol', and Faure sequences are all different types of (t, s)-sequences. Niederreiter proposes a special construction [126] of a (t, s)-sequence using irreducible polynomials which yields a smaller discrepancy bound than either the Sobol' or Faure sequences. This special construction of the (t, s)-sequence is simply referred to as the Niederreiter sequence in base q, throughout this investigation.

Before one can define a (t, s)-sequence in base q, one must first define an elementary interval and a (t, m, s)-net. An *elementary interval in base q* is a subinterval E of $[0, 1)^s$ with the form

$$E = \prod_{i=1}^{s} \left[a_i q^{-d_i}, (a_i + 1) q^{-d_i} \right), \tag{F.1}$$

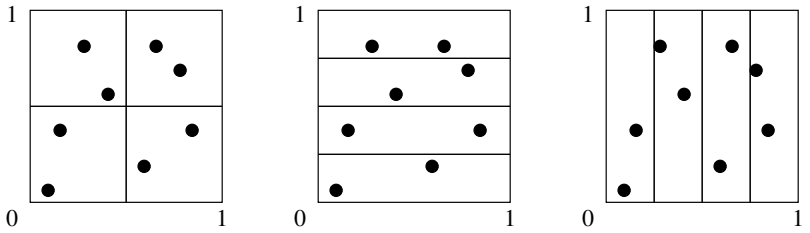

Figure F.1: Example of a (t, m, s)-net in base 2, where $t = 1$, $m = 3$, and $s = 2$.

where a_i and d_i are non-negative integers and $0 \leq a_i \leq b^{d_i}$ for $1 \leq i \leq s$. Observe from the definition in (F.1) that the volume or content of the subinterval $\lambda_s(E) = q^{-D}$, where $D = \sum_{i=1}^{s} d_i$. Next, let $0 \leq t \leq m$ be integers. A *(t, m, s)-net in base q* is then a point set P of q^m points in $[0, 1)^s$ such that every elementary interval with a volume q^{t-m} contains exactly q^t points from P. For example, the $(1, 3, 2)$-net in base 2 is a point set in two dimensions with 8 points distributed such that every elementary subinterval E with an area $\lambda_2(E) = \frac{1}{4}$ contains exactly two points, as illustrated in Figure F.1. The integer t characterizes the strength of the uniformity condition of the (t, m, s)-net in base q. The smaller the value of t, the greater the number of elementary intervals E, each with volume $\lambda_s(E) = q^{t-m}$, that satisfy the uniformity condition for the distribution of points in the (t, m, s)-net in base q.

A sequence of points $\mathbf{x}_0, \mathbf{x}_1, \ldots \in \overline{I}^s$ is a *(t, s)-sequence in base q* if, for all integers $k \geq 0$ and $m > t$, the point set consisting of $\mathbf{x}_n$ with $kq^m \leq n \leq (k+1)q^m$ is a (t, m, s)-net in base q. In other words, for all $m > t$, every successive block of q^m elements from a (t, s)-sequence in base q is also a (t, m, s)-net in base q. Remarkably, without actually defining the precise location of the sequence elements, Niederreiter [125, 127] proves the asymptotic bound on the star discrepancy of a (t, s)-sequence in base q is given by

$$D_N^* \leq C_s^N \frac{(\log N)^s}{N} + \mathcal{O}\left(q^t N^{-1} (\log N)^{s-1}\right), \tag{F.2}$$

where

$$C_s^N = \begin{cases} \frac{1}{s}\left(\frac{q-1}{2\log q}\right)^s q^t & \text{if } s = 2 \text{ or } q = 2,\ s = 3, 4 \\ \frac{1}{s!} \cdot \frac{q-1}{2\lfloor q/2 \rfloor}\left(\frac{\lfloor q/2 \rfloor}{\log q}\right)^s q^t & \text{otherwise.} \end{cases} \tag{F.3}$$

Note that the operator $\lfloor \cdot \rfloor$ in (F.3) is the floor function. Therefore, the bound in (F.2) implies that every (t, s)-sequence is a low-discrepancy sequence as well.

The special construction of Niederreiter sequence in base q [126] uses irreducible polynomials in $\mathbb{F}_q[x]$. Refer to [96] for a thorough review of finite fields and irreducible polynomials. The uniformity measure t for the (t, s) classification of the Niederreiter sequences in base q is related to the degree of the irreducible polynomials used to generate each dimension; specifically,

$$t = \sum_{i=1}^{s} \big(\deg(p_i(x)) - 1\big),$$

where $p_i(x) \in \mathbb{F}_q[x]$ is the irreducible polynomial used to generate the i^{th} dimension of the sequence, and $p_1(x), \ldots, p_s(x)$ are distinct.. Thus, the most uniform-(t, s) sequence is constructed using the irreducible polynomials with the lowest possible degree, which is similar to the Sobol' sequence.

The most prominent difference between the Niederreiter and Sobol' sequences is the choice of generating polynomials. The Sobol' sequence is constructed from primitive polynomials in $\mathbb{F}_2[x]$, while the Niederreiter sequence in base q uses irreducible polynomials in $\mathbb{F}_q[x]$, where q is a prime power. Every primitive polynomial is irreducible, but not every irreducible polynomial is necessarily primitive; as a consequence, given an integer $d \geq 1$, there are more irreducible polynomials than primitive polynomials in $\mathbb{F}_2[x]$ with a degree less than or equal to d. Therefore, the Niederreiter sequence in base 2 has a smaller parameter t than the Sobol' sequence of the same dimension.[1]

[1]Because the non-primitive polynomial $p(x) = x$ can still be used to construct the Sobol' se-

The Niederreiter sequence relies on the formal Laurent series to obtain something equivalent to the inverse of the irreducible polynomials used to generate each sequence dimension. For a specific generating polynomial $p(x)$, this inverse is used in a manner analogous to the van der Corput sequence to generate a reflection of the sequence number n around the decimal point to calculate the n^{th} element of the sequence. However, the sequence element number n is not represented with integer digits, but with polynomials from the residue class $F_q[x]/p(x)$ to create a base $p(x)$ representation of n. Irreducible polynomials are similar to primes in that they have no proper divisors in their respective rings other than one and themselves (or scalar multiples of these two possibilities). Since each dimension of the Niederreiter sequence is generated from a distinct irreducible polynomial, the construction is basically an extension of the Halton, except with a set of distinct irreducible polynomial bases instead of primes. While on the surface the construction details may not appear to be similar, the Niederreiter sequence in base 2 is, in fact, closely related to the Sobol' sequence. In particular, the calculation of the formal Laurent series over the irreducible polynomials in $\mathbb{F}_2[x]$ for the Niederreiter sequence is essentially the same as the linear recurrence equations used to produce the direction numbers for the Sobol' sequence.

The general construction of the Niederreiter sequence in base q is initially presented here for only one dimension in order to simplify the required notation. The dimensions of the Niederreiter sequence are generated independently; thus, the construction process that follows for the one dimensional sequence is repeated as needed for a multi-dimensional sequence. Since each dimension is generated by a distinct irreducible polynomial over a finite field, let $p(x) \in \mathbb{F}_q[x]$ represent the irreducible

quence, the value t for the Niederreiter sequence in base 2 is not strictly less than the Sobol' sequence unless $t \geq 8$, see Table F.2.

polynomial under consideration for this one dimensional example of a Niederreiter sequence in base q. Tables of irreducible polynomials over finite fields are found in [96], or they may be produced by a simple extension of the Sieve of Eratosthenes algorithm [78] from the integers to $\mathbb{F}_q[x]$. The construction in this example is limited to the first q^k sequence elements, where k is some positive integer, in order to eliminate the need for infinite summations. Furthermore, a matrix-vector notation similar to that used in the construction of the Faure sequence (see Appendix E) is adopted for the Niederreiter sequence as well.

For $0 \leq n \leq q^k - 1$, let $x_n \in [0, 1)$ denote the n^{th} element of a one dimensional Niederreiter sequence in base q generated by the irreducible polynomial $p(x) \in \mathbb{F}_q[x]$, where q is a prime power. Using a finite analogue of the base b representation vector defined in (A.1) for the van der Corput sequence, let $\vec{\xi_q}(n) = (\xi_{1,q}(n), \ldots, \xi_{k,q}(n))$ denote the base q representation of an integer n. Next define the matrix $A = [a_{ij}] \in \mathbb{F}_q^{k \times k}$ such that the i^{th} row of coefficients $(a_{i1}, \ldots, a_{ik})$ is defined in terms of the formal Laurent series $\mathbb{F}_q((x^{-1}))$ of the following polynomial quotient:

$$\frac{x^{\mathrm{mod}(i-1,d)}}{p(x)^{\lceil i/d \rceil}} = a_{i1}x^{-1} + a_{i2}x^{-2} + \cdots + a_{ik}x^{-k} + \cdots \in F_q((x^{-1})) \quad \text{for } 1 \leq i, j \leq k, \tag{F.4}$$

where $\lceil \cdot \rceil$ denotes the ceiling function. The calculation of the formal Laurent series of a polynomial coefficient is reviewed in [127]. Let the vector $\mathbf{y}_n = (y_{1,n}, \ldots, y_{k,n})$ denote the result from the matrix-vector multiplication given by

$$\mathbf{y}_n = A\vec{\xi_q}(n). \tag{F.5}$$

It is important to note that the addition and multiplication operation in (F.5) are performed over the finite field $\mathbb{F}_q$. Finally, the vector $\mathbf{y}_n$ in (F.5) is used to calculate the n^{th} element of the Niederreiter sequence in base q in the same manner as the

Faure sequence; that is,

$$x_n = \sum_{i=1}^{k} y_{i,n} q^{-i}, \tag{F.6}$$

where the calculations are performed over the real numbers (*i.e.* standard arithmetic).

As an example, consider a three dimensional Niederreiter sequence in base 3, with a maximum number of digits $k = 4$. Suppose that the i^{th} dimension of this Niederreiter sequence is generated by the irreducible polynomials $p_i(x) \in \mathbb{F}_3[x]$, where $p_1(x) = x+1$, $p_2(x) = x^2+1$, and $p_3(x) = x^2+x+2$. In order to generate the matrix $A^{(m)}$ in each dimension $m = 1, 2, 3$, the four leading order terms of the following formal Laurent series $\mathbb{F}_3((x^{-1}))$ are needed. Note that a superscript representing the sequence dimension is added to the A matrix (F.4) and the vector $\mathbf{y}_n$ (F.5). For $p_1(x) = x + 1$,

$$\begin{aligned}
\text{row 1:} \quad \tfrac{1}{p_1(x)} &= x^{-1} + 2x^{-2} + x^{-3} + 2x^{-4} + \cdots \\
\text{row 2:} \quad \tfrac{1}{p_1(x)^2} &= x^{-2} + x^{-3} + x^{-5} + \cdots \\
\text{row 3:} \quad \tfrac{1}{p_1(x)^3} &= x^{-3} + 2x^{-6} + \cdots \\
\text{row 4:} \quad \tfrac{1}{p_1(x)^4} &= x^{-4} + \cdots
\end{aligned}$$

which yields the following matrix for the first dimension:

$$A^{(1)} = \begin{pmatrix} 1 & 2 & 1 & 2 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{F.7}$$

For $p_2(x) = x^2 + 1$,

$$\begin{aligned}
\text{row 1:} \quad \tfrac{1}{p_2(x)} &= x^{-2} + 2x^{-4} + \cdots \\
\text{row 2:} \quad \tfrac{x}{p_2(x)} &= x^{-1} + 2x^{-3} + x^{-5} + \cdots \\
\text{row 3:} \quad \tfrac{1}{p_2(x)^2} &= x^{-4} + \cdots \\
\text{row 4:} \quad \tfrac{x}{p_2(x)^2} &= x^{-3} + x^{-5} \cdots
\end{aligned}$$

which yields the following matrix for the second dimension:

$$A^{(2)} = \begin{pmatrix} 0 & 1 & 0 & 2 \\ 1 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}. \tag{F.8}$$

For $p_3(x) = x^2 + x + 2$,

$$\begin{aligned} \text{row 1:} \quad \tfrac{1}{p_3(x)} &= x^{-2} + 2x^{-3} + 2x^{-4} + \cdots \\ \text{row 2:} \quad \tfrac{x}{p_3(x)} &= x^{-1} + 2x^{-2} + 2x^{-3} + 2x^{-5} \cdots \\ \text{row 3:} \quad \tfrac{1}{p_3(x)^2} &= x^{-4} + \cdots \\ \text{row 4:} \quad \tfrac{x}{p_3(x)^2} &= x^{-3} + x^{-4} \cdots \end{aligned}$$

which yields the following matrix for the third dimension:

$$A^{(3)} = \begin{pmatrix} 0 & 1 & 2 & 2 \\ 1 & 2 & 2 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}. \tag{F.9}$$

Consider again, for the moment, the general case of a Niederreiter sequence in base q, where the maximum number of digits $k \geq 1$ and the degree of the irreducible polynomial $p(x) \in \mathbb{F}_q[x]$ is $d \geq 1$. In this general case, the calculation of the A matrix in (F.4) only requires the first $d + k$ coefficients of the formal Laurent series for $\frac{1}{p(x)^r}$, for $1 \leq r \leq \lceil \frac{k}{d} \rceil$. Once the coefficients from the formal Laurent series of $\frac{1}{p(x)^r}$ are known, it is possible to obtain the coefficients for the formal Laurent series of any polynomial quotient of the form $\frac{x^j}{p(x)^r}$ by simply shifting the coefficients to the right j places.

Now that the A matrices (F.4) have been established for this example of the Niederreiter sequence in base 3, the actual elements $\mathbf{x}(n) = (x_1(n), x_2(n), x_3(n))$ of

the sequence can be calculated for $0 \leq n \leq 3^4 - 1$. To illustrate the actual process of generating this sequence, consider the calculation for the 13^{th} element of the sequence $\mathbf{x}_{13}$. First one must determine the base 3 representation of 13; that is,

$$\vec{\xi}_3(13) = (1, 1, 1, 0)^T.$$

Using the previously defined $A^{(m)}$ matrices in (F.7-F.9), the matrix-vector multiplication in (F.5) is then applied for $m = 1, 2, 3$ in order to calculate the vectors $\mathbf{y}_{13}^{(m)}$, which yields

$$\begin{aligned} \mathbf{y}_{13}^{(1)} &= (1, 1, 1, 0)^T \\ \mathbf{y}_{13}^{(2)} &= (0, 0, 1, 0)^T \\ \mathbf{y}_{13}^{(3)} &= (1, 2, 1, 0)^T. \end{aligned}$$

Finally, this set of vectors is used in (F.6) to generate each coordinate of the 13^{th} element of the Faure sequence; specifically,

$$\mathbf{x}_{13} = (\tfrac{16}{27}, \tfrac{28}{81}, \tfrac{19}{81}).$$

For further reference, the calculation of the first 16 points of this Niederreiter sequence in base 3 is provided in Table E.1.

The Niederreiter sequence offers greater control over the bounding constant for its discrepancy bound (F.3) than the Sobol' and Faure sequences. In particular, if an s-dimensional Niederreiter sequence is constructed in base 2, then the discrepancy bound is less than that of the Sobol' sequence (when $s \geq 8$). Further, if the Niederreiter sequence is constructed in a prime power base $q \geq s$, then the discrepancy bound is less than or equal to that of the Faure sequence. When the base of the Niederreiter sequence is greater than or equal to the dimension of the sequence, all the polynomials used in its construction are of degree one. Thus, this Niederreiter sequence has

n	$x_1(n)$	$x_2(n)$	$x_3(n)$
0	0.0000	0.0000	0.0000
1	0.3333	0.1111	0.1111
2	0.6667	0.2222	0.2222
3	0.7778	0.3333	0.5556
4	0.1111	0.4444	0.3333
5	0.4444	0.5556	0.4444
6	0.5556	0.6667	0.7778
7	0.8889	0.7778	0.8889
8	0.2222	0.8889	0.6667
9	0.4814	0.2346	0.9012
10	0.8148	0.0123	0.6790
11	0.1481	0.1235	0.7901
12	0.2593	0.5679	0.1235
13	0.5926	0.3457	0.2346
14	0.9259	0.4568	0.0123
15	0.7037	0.9012	0.3457

Table F.1: The first 16 points constructed for a three dimensional Niederreiter sequence in base 3. The following irreducible polynomials are used to generate each dimension: $p_1(x) = x+1$, $p_2(x) = x^2+1$ and $p_3(x) = x^2+x+2$ in $\mathbb{F}_3[x]$.

a value $t = 0$, using the (t, s) classification. As mentioned earlier, $(0, s)$-sequences are considered the most uniform; in fact, it can be shown that the implied constant C_s^N in (F.3) tends to zero as $s \to \infty$ in this case. It is interesting to note that the s-dimensional Niederreiter sequence in base q when $q \geq s$ is prime is the same as the Faure sequence in base q, except for a re-ordering of the sequence dimensions. Based on the asymptotic performance in (F.2), it is natural to assume that $(0, s)$-type of Niederreiter sequence would be the best option for QMC integration. In practice, however, the Niederreiter sequence in base 2 offers the best performance because the computational cost of generating the sequence is orders of magnitude smaller than that of the $(0, s)$-sequence in base $q > 2$.

For the Niederreiter sequence in base 2, the matrix-vector multiplication is reduced to a vector product by concatenating the ones and zeros in each column of the matrix transform A into a single integer. The base 2 calculation can be further

simplified to a single operation by adopting the Gray code modification that Antonov and Saleev [2] propose for the Sobol' sequence. In contrast, the Niederreiter sequence in a general base $q > 2$ requires the full matrix-vector multiplication to construct each dimension. When the sequence base p is a prime number greater than 2, all the additions and multiplications needed for the matrix-vector multiplication must be performed modulo p which adds an additional division operation for each row of the matrix. Furthermore, when the sequence base is a prime power p^r ($r > 1$), there is no simple means to perform the addition and multiplication operations over the finite field $\mathbb{F}_{p^r}$. In this case, the full operator tables for addition and multiplication on $\mathbb{F}_{p^r}$ must be stored in memory and accessed for each operation in the matrix-vector multiplication. More details on the computational cost associated with the Niederreiter sequence in base 2 and the $(0, s)$-type sequences (*e.g.* the Faure sequence) are given in Section 4.3.

The performance advantages of the Niederreiter sequence in base 2 leads to a natural question: "if the (t, s)-sequence in base 2 is preferable in practice and lower values of t result in a more uniform sequence, what is the lowest value of t that can be achieved for a given s?" For the Sobol' sequence, the value of $t = t(s)$ grows as the degree increases of the first s primitive polynomials in $\mathbb{F}_2[x]$ (when ordered by lowest degree). Similarly, for the Niederreiter sequence in base 2, the value of $t = t(s)$ grows as the degree increases of the first s irreducible polynomials $\mathbb{F}_2[x]$. However, since not every irreducible polynomial is a primitive polynomial, there is generally more irreducible polynomials in $\mathbb{F}_2[x]$ of a specific degree. Thus, the value of $t(s)$ for the Niederreiter sequence grows slower than the Sobol' sequence, yielding a smaller bounding constant (F.3) in the discrepancy bound.

Niederreiter and Xing [130] extend this concept further by constructing (t, s)-

s	Sobol' (1967)	Nied (1988)	NX (1996)
1	0	0	0
2	0	0	0
3	1	1	1
4	3	3	1
5	5	5	2
6	8	8	3
7	11	11	4
8	15	14	5
9	19	18	6
10	23	22	8
11	27	26	9
12	31	30	10
13	35	34	11
14	40	38	13
15	45	43	15
16	50	48	15
17	55	53	18
18	60	58	19
19	65	63	19
20	71	68	21

Table F.2: The value $t = t(s)$, for $1 \leq s \leq 20$, of different (t, s) sequences in base 2. Originally from [129].

sequences using global function fields with many rational places. The same construction ideas behind low-discrepancy sequences have evolved from using prime numbers (Halton) to using finite field polynomials (Sobol' and Niederreiter) and then to using global function fields (Niederreiter and Xing). Each step in this evolution relies on more abstract algebraic and number theoretic concepts, thereby allowing for greater control over the asymptotic convergence of the star discrepancy of the sequence. The (t, s)-sequence in base 2 proposed by Niederreiter and Xing has the slowest growth of $t(s)$ currently known. A comparison of $t(s)$ for the base 2 sequences of Sobol', Niederreiter (Nied), and Niederreiter and Xing (NX) is given in [129], which is reproduced here in Table F.2.

While Niederreiter and Xing proved in [130] the existence of a (t, s)-sequence in base 2 with the values of $t(s)$ in Table F.2, there was no explicit construction provided

for the sequence. Over four years elapsed before the first actual implementations [144] of the Niederreiter and Xing sequence were performed. Many of the necessary procedures in this new construction are still very active areas in algebra research. Thus, the algorithms for the Niederreiter and Xing sequence will continue to evolve for quite some time.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] M. Abramowitz and I. A. Stegun, editors. *Handbook of Mathematical Functions*. Dover, New York, 9^{th} edition, 1972.

[2] I. A. Antonov and V. M. Saleev. An economic method of computing LP_τ − sequences. *USSR Computational Mathematics and Mathematical Physics*, 19:252–256, 1979. (*translated from Russian*).

[3] G. Arfken. *Mathematical Methods for Physicists*. Academic Press, New York, 1968.

[4] E. B. Arkilic, K. S. Breuer, and M. A. Schmidt. Gaseous flow in microchannels. In J. Harvey and G. Lord, editors, 19^{th} *International Symposium on Rarefied Gas Dynamics*, pages 347–353, Oxford, U. K., 1994.

[5] E. B. Arkilic, K. S. Breuer, and M. A. Schmidt. Mass flow and tangential momentum accommodation in silicon micromachined channels. *The Journal of Fluid Mechanics*, 437:29–43, 2001.

[6] E. B. Arkilic, M. A. Schmidt, and K. S. Breuer. TMAC measurement in silicon micromachined channels. In Ching Shen, editor, 20^{th} *International Symposium on Rarefied Gas Dynamics*, pages 983–988, Beijing, China, 1996.

[7] K. E. Atkinson. *A Survey of Numerical Methods for the Solution of Fredholm Integral Equations of the Second Kind*. SIAM, Philadelphia, 1976.

[8] H. Babovsky, F. Gropengießer, H. Neunzert, J. Struckmeier, and B. Wiesen. Low-discrepancy method for the Boltzmann equation. In E. P. Muntz, D. P. Weaver, and D. H. Campbell, editors, 16^{th} *International Symposium on Rarefied Gas Dynamics*, pages 85–99, Pasadena, CA, 1988.

[9] H. Babovsky, F. Gropengießer, H. Neunzert, J. Struckmeier, and B. Wiesen. Application of well-distributed sequences to the numerical simulation of the Boltzmann equation. *Journal of Computational and Applied Mathematics*, 31(1):15–22, 1990.

[10] H. Babovsky and R. Illner. A convergence proof for Nanbu's simulation method for the full Boltzmann equation. *SIAM Journal of Numerical Analysis*, 26(1):45–65, 1989.

[11] P. Bahukudumbi and A. Beskok. A phenomenological lubrication model for the entire Knudsen regime. *Journal of Micromechanics and Microengineering*, 13:873–884, 2003.

[12] P. Bahukudumbi, J. H. Park, and A. Beskok. A unified engineering model for steady and quasi-steady shear driven gas microflows. *Microscale Thermophysical Engineering*, 7(4):291–315, 2003.

[13] L. Bernstein. *The Jacobi-Perron Algorithm; its theory and application.* Springer-Verlag, New York, 1971.

[14] A. S. Besicovitch. On the linear independence of fractional powers of integers. *Journal of the London Mathematical Society*, 15:3–6, 1940.

[15] A. Beskok and G. Karniadakis. A model for flows in channels, pipes, and ducts at micro and nano scales. *Microscale Thermophysical Engineering*, 3(1):43–77, 1999.

[16] G. A. Bird. *Molecular Gas Dynamics and the Direct Simulation of Gas Flows.* Clarendon Press, Oxford, 1994.

[17] M. Bordeau and A. Pitre. Tables of good lattices in four and five dimensions. *Numerische Mathematik*, 47:39–43, 1985.

[18] P. Bratley and B. L. Fox. Algorithm 659: Implementing Sobol' quasirandom sequence generator. *ACM Transactions on Mathematical Software*, 14(1):88–100, 1988.

[19] P. Bratley, B. L. Fox, and H. Niederreiter. Implementation and test of low-discrepancy sequences. *ACM Transactions on Modeling and Computer Simulation*, 2(3):195–213, 7 1992.

[20] R. L. Burden and J. D. Faires. *Numerical Analysis.* Brooks/Cole, Pacific Grove, CA, 7^{th} edition, 2001.

[21] D. Burnett. The distribution of molecular velocities and the mean motion in a non-uniform gas. *Proceedings of the London Mathematical Society*, 40:382–435, 1935.

[22] D. Burnett. The distribution of velocities in a slightly non-uniform gas. *Proceedings of the London Mathematical Society*, 39:385–430, 1935.

[23] R. E. Caflisch and B. Moskowitz. Modified Monte-Carlo methods using quasi-random sequences. In H. Niederreiter and P. Shiue, editors, *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, pages 1–16, Las Vegas, NV, 1994.

[24] C. Cai, I. D. Boyd, and J. Fang. Direct simulation methods for low-speed microchannel flows. *Journal of Thermophysics and Heat Transfer*, 14(3):368–378, 2000.

[25] C. Cercignani. *Mathematical Methods in Kinetic Theory*. Plenum Press, New York, 1969.

[26] C. Cercignani and M. Lampis. Kinetic models for gas-surface interactions. *Transport Theory and Statistical Physics*, 1(2):101–114, 1971.

[27] I. Chakraborty, W. C. Tang, D. P. Bame, and T. K. Tang. MEMS micro-valve for space applications. *Sensors and Actuators*, 83:188–193, 2000.

[28] S. Chapman and T. G. Cowling. *The Mathematical Theory of Non-Uniform Gases*. Cambridge University Press, Cambridge, U.K., 1952.

[29] C. Chen and J. G. Santiago. A planar electroosmotic micropump. *Journal of Microelectromechanical Systems*, 11(6):672–683, 2002.

[30] P. Clausing. Über die Strömung sehr verdünnter Gase durch Röhren von beliebuger Länge. *Annalen der Physik*, 404(8):961–989, 1932. (*in German*).

[31] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. McGraw-Hill, New York, 2^{nd} edition, 2001.

[32] RAND Corportation. *A Million Random Digits with 100,000 Normal Deviates*. RAND Corportation, Pittsburgh, 2001.

[33] I. Coulibaly and C. Lécot. Monte Carlo and quasi-Monte Carlo algorithms for a linear integro-differential equation. In H. Niederreiter, P. Hellekalek, G. Larcher, and P. Zinterhof, editors, 2^{nd} *International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, pages 176–188, Salzburg, Austria, 1996.

[34] R. Courant and D. Hilbert. *Methods of Mathematical Physics*, volume 1. Interscience Publishers, New York, 1953.

[35] R. Cranley and T. N. L. Patterson. Randomization of number theoretic methods for multiple integration. *SIAM Journal on Numerical Analysis*, 13(6):904–914, 1976.

[36] D. H. Davis. Monte Carlo calculation of molecular flow rates through a cylindrical elbow and pipes of other shapes. *Journal of Applied Physics*, 31(7):1169–1176, 1960.

[37] M. de Berg, M. van Kreveld, M. Overmars, and O. Schwarzkopf. *Computational Geometry: Algorithms and Applications*. Springer-Verlag, New York, 1998.

[38] J. H. de Boer. *The Dynamical Character of Adsorption*. Clarendon Press, Oxford, 1953.

[39] R. G. Deissler. An analysis of second-order slip flow and temperature-jump boundary conditions for rarefied gases. *International Journal of Heat and Mass Transfer*, 7:681–694, 1964.

[40] L. M. Delves and J. L. Mohamed. *Computational Methods for Integral Equations*. Cambridge University Press, New York, 1985.

[41] L. Devroye. Binary search trees based on Weyl and Lehmer sequences. In H. Niederreiter, P. Hellekalek, G. Larcher, and P. Zinterhof, editors, 2^{nd} *International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, pages 40–65, Salzburg, Austria, 1996.

[42] J. Dick and F. Y. Kuo. Constructing good lattice rules with millions of points. In H. Niederreiter, editor, 5^{th} *International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, pages 181–198, Singapore, 2002.

[43] S. Dushman. Recent advances in the production and measurement of high vacua. *Journal of the Franklin Institute*, 211(6):689–750, 1931.

[44] J. Fan and C. Shen. Statistical simulation of low-speed rarefied gas flows. *Journal of Computational Physics*, 167:393–412, 2001.

[45] H. Faure. Discrépance de suites associées à un système de numération (en dimension un). *Bulletin de la Société Mathématique de France*, 109:143–182, 1981. (*in French*).

[46] H. Faure. Discrépance de suites associées à un système de numération (en dimension s). *Acta Arithmetica*, 41:337–351, 1982. (*in French*).

[47] G. S. Fishman. *Monte Carlo: concepts, algorithms, and applications*. Springer-Verlag, New York, 1996.

[48] B. L. Fox. Algorithm 647: Implementation and relative efficiency of quasirandom sequence generators. *ACM Transactions on Mathematical Software*, 12(4):362–376, 1986.

[49] R. G. J. Fraser. *Molecular Rays,*. Cambridge University Press, Cambridge, 1931.

[50] S. H. Friedberg, A. J. Insel, and L. E. Spence. *Linear Algebra*. Prentice Hall, Upper Saddle River, NJ, 4^{th} edition, 2003.

[51] M. Gad-el Hak. The fluid mechanics of microdevices - the Freeman Scholar Lecture. *Journal of Fluids Engineering*, 121:5–33, 3 1999.

[52] M. Gardner. *Knotted Doughnuts and Other Mathematical Entertainments*. W. H. Freeman and Company, New York, 1986.

[53] A. Gerlach et al. Microfabrication of single-use plastic microfluidic devices for high-throughput screening and DNA analysis. *Microsystem Technology*, 7:265–268, 2002.

[54] T. I. Gombosi. *Gas Kinetic Theory*. Cambridge University Press, New York, 1994.

[55] C. P. T. Groth, P. L. Roe, T. I. Gombosi, and S. L. Brown. On the non-stationary wave structure of a 35-moment closure for rarefied gas dynamics, AIAA-1995-2312. In 26^{th} *AIAA Fluid Dynamics Conference*, San Diego, CA, 1995.

[56] J. H. Halton. On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. *Numerische Mathematik*, 2:84–90, 1960.

[57] J. H. Halton and G. B. Smith. Algorithm 247 radical-inverse quasi-random point sequence. *Communications of the ACM*, 7(12):701–702, 1964.

[58] J. M. Hammersley and D. C. Handscomb. *Monte Carlo Methods*. John Wiley & Sons, New York, 1964.

[59] S. Heinrich. Efficient algorithms for computing the L_2-discrepancy. *Mathematics of Computation*, 65:1621–1633, 1996.

[60] S. Heinrich and A. Keller. Quasi-Monte Carlo methods in computer graphics, part I: The QMC-buffer. Technical report, Universität Kaiserslautern, Kaiserslautern, Germany, 1994.

[61] O. Henry. *The complete works of O. Henry*. Doubleday, Garden City, NY, 1960.

[62] F. J. Hickernell. A generalized discrepancy and quadrature error bound. *Mathematics of Computation*, 67:299–322, 1998.

[63] C. Hirsch. *Numerical Computation of Internal and External Flows*, volume 1–2. John Wiley & Sons, New York, 2000.

[64] D. L. Hitt, C. M. Zakrzwski, and M. A. Thomas. MEMS-based satellite micropropulsion via catalyzed hydrogen peroxide decomposition. *Smart Materials and Structures*, 10:1163–1175, 2001.

[65] J. A. Hittinger. *Foundations for the Generalization of the Godunov Method to Hyperbolic Systems with Stiff Relaxation Source Terms*. PhD thesis, The University of Michigan, 2000.

[66] E. Hlawka. Funktionen von beschränkter Variation in der Theorie der Gleichverteilung. *Annali di Matematica Pura ed Applicata*, 54:325–333, 1961. (*in German*).

[67] C. M. Ho and Y. C. Tai. Micro-Electro-Mechanical-Systems (MEMS) and fluid flows. *Annual Review of Fluid Mechanics*, 30:579–612, 1998.

[68] F. James. Monte Carlo theory and practice. *Reports on Progress in Physics*, 43:1145–1189, 1980.

[69] G. Karniadakis and A. Beskok. *Micro Flows*. Springer-Verlag, New York, 2002.

[70] G. Kedem and S. K. Zaremba. A table of good lattice points in three dimensions. *Numerische Mathematik*, 23:175–180, 1974.

[71] A. Keller. A quasi-Monte Carlo algorithm for the global illumination problem in the radiosity setting. In H. Niederreiter and P. Shiue, editors, *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, pages 239–251, Las Vegas, NV, 1994.

[72] A. Keller. The quasi-Monte Carlo walk. In H. Niederreiter, P. Hellekalek, G. Larcher, and P. Zinterhof, editors, 2^{nd} *International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, pages 277–291, Salzburg, Austria, 1996.

[73] E. H. Kennard. *Kinetic Theory of Gases*. McGraw-Hill, New York, 1938.

[74] A. Kersch, W. J. Morokoff, and A. Schuster. Radiative heat transfer with quasi-Monte Carlo methods. *Transport Theory and Statistical Physics*, 23(7):1001–1021, 1994.

[75] A. Ya. Khinchin. *Continued Fractions*. Dover, New York, 1997.

[76] M. Knudsen. Die Gesetze der Molekularströmung und der inneren Reibungßtrömung der Gase durch Röhren. *Annalen der Physik*, 333(1):75–130, 1909. (*in German*).

[77] M. Knudsen. Die Molekularströmung der Gase durch öffnungen und die Effusion. *Annalen der Physik*, 333(5):999–1016, 1909. (*in German*).

[78] D. E. Knuth. *The Art of Computer Programming*, volume 2. Addison-Wesley, New York, 3^{rd} edition, 1998.

[79] D. E. Knuth. *The Art of Computer Programming*, volume 3. Addison-Wesley, New York, 3^{rd} edition, 1998.

[80] M. N. Kogan. *Rarefied Gas Dynamics*. Plenum Press, New York, 1969. (*translated from Russian*).

[81] J. F. Koksma. Een algemeene stelling uit de theorie der gelijkmatige verdeeling modulo 1. *Mathematica (Zutphen) B*, 11:7–11, 1942/1943. (*in Dutch*).

[82] N. M. Korobov. The approximate computation of multiple integrals. *Doklady Akademii Nauk SSSR*, 124:1207–1210, 1959. (*in Russian*).

[83] N. M. Korobov. Properties and calculation of optimum coefficients. *Soviet Mathematics Doklady*, 1:696–701, 1960. (*translated from Russian*).

[84] M. Krook and T. T. Wu. Exact solutions of the Boltzmann equation. *The Physics of Fluids*, 20(10):1589–1595, 1977.

[85] L. Kuipers and H. Niederreiter. *Uniform Distributions of Sequences*. John Wiley & Sons, New York, 1974.

[86] C. Lécot. A direct simulation Monte Carlo scheme and uniformly distributed sequences for solving the Boltzmann equation. *Computing*, 41:41–57, 1989.

[87] C. Lécot. Low discrepancy sequences for solving the Boltzmann equation. *Journal of Computational and Applied Mathematics*, 25:237–249, 1989.

[88] C. Lécot. A quasi-Monte Carlo method for the Boltzmann equation. *Mathematics of Computation*, 56:621–644, 4 1991.

[89] C. Lécot and I. Coulibaly. A quasi-Monte Carlo scheme using nets for a linear Boltzmann equation. *SIAM Journal of Numerical Analysis*, 35(1):51–70, 2 1998.

[90] P. L'Ecuyer. Uniform random number generators: A review. In S. Andradóttir, K. Healy, D. Withers, and B. Nelson, editors, *Proceedings of the 1997 Winter Simulation Conference*, pages 127–134, 1997.

[91] P. L'Ecuyer. Uniform random number generators. In D. Medeiros, E. Watson, J. Carson, and M. Manivannan, editors, *Proceedings of the 1998 Winter Simulation Conference*, pages 97–104, 1998.

[92] L. Lees. Kinetic theory description of rarefied gas flows. *Journal of the Society for Industrial and Applied Mathematics*, 13(1):278–311, 1965.

[93] K. Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics*, 2:164–168, 1944.

[94] C. D. Levermore. Moment closure hierarchies for kinetic theories. *Journal of Statistical Physics*, 83(5/6):1021–1065, 1996.

[95] C. D. Levermore and W. J. Morokoff. The Gaussian moment closure for gas dynamics. *SIAM Journal on Applied Mathematics*, 59(1):72–96, 1998.

[96] R. Lidl and H. Niederreiter. *Introduction to Finite Fields and their Applications*. University Press, Cambridge, 1994.

[97] M. Livio. *The Golden Ratio: The Story of Phi, the World's Most Astonishing Number*. Broadway Books, New York, 2002.

[98] L. B. Loeb. *Kinetic Theory of Gases*. McGraw-Hill, New York, 1934.

[99] R. G. Lord. Tangential momentum accommodation coefficients of rare gases on polycrystalline metal surfaces. In J. L. Potter, editor, 10^{th} *International Symposium on Rarefied Gas Dynamics*, pages 531–538, Aspen, CO, 1976.

[100] R. G. Lord. Some extensions to the Cercignani-Lampis gas-surface scattering kernel. *Physics of Fluids A*, 3(11):706–710, 1991.

[101] R. G. Lord. Some further extensions to the Cercignani-Lampis gas-surface interaction model. *Physics of Fluids*, 7(5):1159–1161, 1995.

[102] H. A. Lorentz. *Lectures on Theoretical Physics*, volume 1. Macmillan, London, 1927.

[103] S. K. Loyalka and K. A. Hickey. Velocity slip and defect: hard sphere gas. *Physics of Fluids A*, 1(3):612–614, 1989.

[104] J. N. Lyness and T. S. Sørevik. A search program for finding optimal integration lattices. *Computing*, 47:103–120, 1991.

[105] E. B. Magrab et al. *An Engineer's Gide to MATLAB*. Prentice Hall, Upper Saddle River, NJ, 2000.

[106] D. Maisonneuve. Recherche et utilisation des bons treillis, programming et resultats numeriques. In S. K. Zaremba, editor, *Applications of Number Theory to Numerical Analysis*, pages 121–201. Academic Press, New York, 1972. (*in French*).

[107] D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.

[108] M. J. Martin, K. J. Scavazze, I. D. Boyd, and L. P. Bernal. Design of a low-turbulence, low-pressure wind-tunnel for micro-aerodynamics. *Journal of Fluids Engineering - Transactions of the ASME*, 128(5):1045–1052, 2006.

[109] J. C. Maxwell. *The Scientific Papers of James Clerk Maxwell*, volume 2. Dover, New York, 1952.

[110] M. J. McNenly and I. D. Boyd. Numerical simulation of particle flow conductance in a duct under free-molecular conditions. In M. Capitelli, editor, *Rarefied Gas Dynamics*, volume 762, pages 202–207, Melville, NY, 2005. AIP Conference Proceedings.

[111] M. J. McNenly, M. A. Gallis, and I. D. Boyd. Slip model performance for microscale gas flows, AIAA-03-4050. In *The* 36^{th} *AIAA Thermophysics Conference*, Orlando, FL, 2003.

[112] M. J. McNenly, M. A. Gallis, and I. D. Boyd. Empirical slip and viscosity model performance for microscale gas flow. *International Journal for Numerical Methods in Fluids*, 49:1169–1191, 2005.

[113] R. A. Millikan. Coefficients of slip in gases and the law of reflection of molecules from the surfaces of solids and liquids. *The Physical Review*, 21(3):217–238, 1923.

[114] W. J. Morokoff. Generating quasi-random paths for stochastic processes. *SIAM Review*, 40(4):765–788, 1998.

[115] W. J. Morokoff and R. E. Caflisch. A quasi-Monte Carlo approach to particle simulation of the heat equation. *SIAM Journal on Numerical Analysis*, 30(6):1558–1573, 1993.

[116] W. J. Morokoff and R. E. Caflisch. Quasi-random sequences and their discrepancies. *SIAM Journal on Scientific Computing*, 15(6):1251–1279, 1994.

[117] W. J. Morokoff and R. E. Caflisch. Quasi-Monte Carlo integration. *Journal of Computational Physics*, 122:218–230, 1995.

[118] W. J. Morokoff and R. E. Caflisch. Quasi-Monte Carlo simulation of random walks in finance. In H. Niederreiter, P. Hellekalek, G. Larcher, and P. Zinterhof, editors, 2^{nd} *International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, pages 340–352, Salzburg, Austria, 1996.

[119] B. Moskowitz. Quasirandom diffusion Monte Carlo. In H. Niederreiter and P. Shiue, editors, *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, pages 278–298, Las Vegas, NV, 1994.

[120] B. Moskowitz and R. E. Caflisch. Smoothness and dimension reduction in quasi-Monte Carlo methods. *Mathematical and Computer Modeling*, 23(8–9):37–54, 1996.

[121] K. Nanbu. Direct simulation scheme derived from the Boltzmann equation. I. monocomponent gases. *Journal of the Physical Society of Japan*, 49(5):2042–2049, 1980.

[122] N. Nguyen, X Huang, and T. K. Chuan. MEMS-micropumps: a review. *Journal of Fluids Engineering - Transactions of the ASME*, 124(6):384–392, 2002.

[123] H. Niederreiter. Methods for estimating discrepancy. In S. K. Zaremba, editor, *Applications of Number Theory to Numerical Analysis*, pages 203–236. Academic Press, New York, 1972.

[124] H. Niederreiter. Application of Diophantine approximations to numerical integration. In C. F. Osgood, editor, *Diophantine Approximation and Its Applications*, pages 129–199. Academic Press, New York, 1973.

[125] H. Niederreiter. Point sets and sequences with small discrepancy. *Monatshefte für Mathematik*, 104:273–337, 1987.

[126] H. Niederreiter. Low-discrepancy and low-dispersion sequences. *Journal of Number Theory*, 30:51–71, 1988.

[127] H. Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods*. SIAM, Philadelphia, 1992.

[128] H. Niederreiter and J. M. Wills. Diskrepanz und Distanz von Maßen bezüglich konvexer und Jordanscher Mengen. *Mathematische Zeitschrift*, 144:125–134, 1975. (*in German*).

[129] H. Niederreiter and C. Xing. The algebraic-geometry approach to low-discrepancy sequences. In H. Niederreiter, P. Hellekalek, G. Larcher, and P. Zinterhof, editors, 2^{nd} *International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, pages 139–160, Salzburg, Austria, 1996.

[130] H. Niederreiter and C. Xing. Low-discrepancy sequences and global function fields with many rational places. *Finite Fields and their Applications*, 2:241–273, 1996.

[131] I. Niven, H. S. Zuckerman, and H. L. Montgomery. *An Introduction to the Theory of Numbers*. John Wiley & Sons, New York, 5^{th} edition, 1991.

[132] M. A. Northrup et al. A MEMS-based miniature DNA analysis system. In 8^{th} *International Conference on Solid-State Sensors and Actuators*, Stockholm, Sweden, 1995.

[133] T. Ohwada, Y. Sone, and K. Aoki. Numerical analysis of the Poiseuille and thermal transpiration flows between two parallel plates on the basis of the Boltzmann equation for hard-sphere molecules. *Physics of Fluids A*, 1(12):2042–2049, 1989.

[134] T. Ohwada, Y. Sone, and K. Aoki. Numerical analysis of the shear and thermal creep flows of a rarefied gas over a plane wall on the basis of the linearized Boltzmann equation for hard-sphere molecules. *Physics of Fluids A*, 1(9):1588–1599, 1989.

[135] A. B. Owen. Randomly permuted (t, m, s)-nets and (t, s)-sequences. In H. Niederreiter and P. Shiue, editors, *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, pages 299–317, Las Vegas, NV, 1994.

[136] A. B. Owen. Multidimensional variation for quasi-Monte Carlo. Technical report, Stanford University Department of Statistics, Stanford, CA, 2004.

[137] V. K. Pamula and R. B. Fair. Detection of nanogram explosive particles with a MEMS sensor. In *SPIE Conference on Detection and Remediation Technologies for Mines and Minelike Targets IV*, Orlando, FL, 1999.

[138] L. S. Pan, G. R. Liu, and K. Y. Lam. Determination of slip coefficient for rarefied gas flows using direct simulation Monte Carlo. *Journal of Micromechanics and Microengineering*, 9(1):89–96, 1999.

[139] L. Pareschi and G. Russo. Time relaxed Monte Carlo methods for the Boltzmann equation. *Transport Theory and Statistical Physics*, 29:415–430, 2000.

[140] L. Pareschi and G. Russo. Time relaxed Monte Carlo methods for the Boltzmann equation. *SIAM Journal on Scientific Computing*, 23(4):1253–1273, 2001.

[141] B. A. Parviz, K. Najafi, M. O. Muller, L. P. Bernal, and P. D. Washabaugh. Electrostatically driven synthetic microjet arrays as a propulsion method for micro flight. *Microsystem Technologies*, 11:1214–1222, 2005.

[142] E. S. Piekos and K. S. Breuer. DSMC modeling of micromechanical devices, AIAA-95-2089. In *The* 30^{th} *AIAA Thermophysics Conference*, San Diego, CA, 1995.

[143] D. A. Pierre. *Optimization Theory with Applications*. Dover, New York, 1986.

[144] G. Pirsic. A software implementation of Niederreiter-Xing sequences. In K. T. Fang, F. J. Hickernell, and H. Niederreiter, editors, 4^{th} *International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, pages 434–445, Hong Kong, 2000.

[145] K. Pong, C. Ho, J. Liu, and Y. Tai. Non-linear pressure distribution in uniform microchannels. In *Application of Microfabrication to Fluid Mechanics*, volume FED-197, pages 51–56. ASME, 1994.

[146] W. H. Press and S. A. Teukolsky. Quasi- (that is, sub-) random numbers. *Computers in Physics*, 3(2):76–79, 1989.

[147] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, New York, 2^{nd} edition, 1992.

[148] R. D. Richtmyer. The evaluation of definite integrals, and a quasi-Monte Carlo method based on the properties of algebraic numbers. Technical report, Los Alamos Scientific Laboratory, Los Alamos, NM, 1951.

[149] K. A. Ross. *Elementary Analysis: The Theory of Calculus*. Springer, New York, 2000.

[150] K. F. Roth. On irregularities of distribution. *Mathematika*, 1:73–79, 1954.

[151] G. Russo, L. Pareschi, S. Trazzi, A. Shevyrin, Ye. Bondar, and M. Ivanov. Comparison between time relaxed Monte Carlo method and majorant frequency scheme methods for the space homogeneous Boltzmann equation. In M. Capitelli, editor, *Rarefied Gas Dynamics*, volume 762, pages 571–576, Melville, NY, 2005. AIP Conference Proceedings.

[152] G. Russo, L. Pareschi, S. Trazzi, A. Shevyrin, Ye. Bondar, and M. Ivanov. Plane Couette flow computations by TRMC and MFS methods. In M. Capitelli, editor, *Rarefied Gas Dynamics*, volume 762, pages 577–582, Melville, NY, 2005. AIP Conference Proceedings.

[153] P. K. Sarkar and M. A. Prasad. A comparative study of pseudo and quasi random sequences for the solution of the integral equations. *Journal of Computational Physics*, 68:66–88, 1987.

[154] S. A. Schaaf and P. L. Chambré. *Flow of Rarefied Gases*. Princeton University Press, Princeton, NJ, 1958.

[155] W. M. Schmidt. Metrical theorems on the fractional parts of sequences. *Transactions of the American Mathematical Society*, 110:493–518, 1964.

[156] T. E. Schwartzentruber and I. D. Boyd. A hybrid particle-continuum method applied to shock waves. *Journal of Computational Physics*, 215:402–416, 2006.

[157] T. E. Schwartzentruber, L. C. Scalabrin, and I. D. Boyd. Hybrid particle-continuum simulations of non-equilibrium hypersonic blunt body flow fields, AIAA-2006-3602. In 9^{th} *AIAA/ASME Joint Thermophysics and Heat Transfer Conference*, San Francisco, CA, 2006.

[158] M. Seidl and E. Steinheil. Measurement of momentum accommodation coefficients on surfaces characterized by auger spectroscopy, SIMS and LEED. In M. Becker and M. Fiebig, editors, 9^{th} *International Symposium on Rarefied Gas Dynamics*, pages E. 9–1, Göttingen, Germany, 1974.

[159] F. Sharipov and D. Kalempa. Velocity slip and temperature jump coefficients for gaseous mixtures. I. viscous slip coefficient. *Physics of Fluids*, 15(6):1800–1806, 2003.

[160] I. H. Sloan and S. Joe. *Lattice Methods for Multiple Integration*. Oxford University Press, Oxford, 1994.

[161] I. M. Sobol'. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics*, 7:86–112, 1967. (*translated from Russian*).

[162] I. M. Sobol'. Uniformly distributed sequences with an additional uniform property. *USSR Computational Mathematics and Mathematical Physics*, 16:236–242, 1976. (*translated from Russian*).

[163] Y. Sone, T. Ohwada, and K. Aoki. Evaporation and condensation on a plane condensed phase: Numerical analysis of the linearized Boltzmann equation for hard-sphere molecules. *Physics of Fluids A*, 1(8):1398–1405, 1989.

[164] Y. Sone, T. Ohwada, and K. Aoki. Temperature jump and Knudsen layer in a rarefied gas over a plane wall: Numerical analysis of the linearized Boltzmann equation for hard-sphere molecules. *Physics of Fluids A*, 1(2):363–370, 1989.

[165] Y. Sone, S. Takata, and T. Ohwada. Numerical analysis of plane Couette flow of a rarefied gas on the basis of the linearized Boltzmann equation for hard-sphere molecules. *European Journal of Mechanics. B. Fluids.*, 9(3):273–288, 1990.

[166] J. Spanier. An analytic approach to variance reduction. *SIAM Journal on Applied Mathematics*, 18(1):172–190, 1970.

[167] J. Spanier. Quasi-Monte Carlo methods for particle transport problems. In H. Niederreiter and P. Shiue, editors, *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, pages 121–148, Las Vegas, NV, 1994.

[168] J. Spanier and L. Li. Quasi-Monte Carlo methods for integral equations. In H. Niederreiter, P. Hellekalek, G. Larcher, and P. Zinterhof, editors, 2^{nd} *International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, pages 398–414, Salzburg, Austria, 1996.

[169] D. Sparks et al. Lab-on-chip system for measuring fluid density and chemical concentration. In 7^{th} *International Conference on Miniaturized Chemical and Biochemical Analysis Systems*, Squaw Valley, CA, 2003.

[170] A. Sreekanth. Slip flow through long circular ducts. In L. Trilling and H.Y. Wachman, editors, 6^{th} *International Symposium on Rarefied Gas Dynamics*, volume 1, pages 667–680, Boston, MA, 1968.

[171] Q. Sun and I. D. Boyd. Numerical simulation of gas flow over micro-scale airfoils, AIAA-2001-3071. In 35^{th} *AIAA Thermophysics Conference*, Anaheim, CA, 2001.

[172] Q. Sun and I. D. Boyd. Theoretical development of the information preservation method for strongly nonequilibrium gas flows, AIAA-2005-4828. In 38^{th} *AIAA Thermophysics Conference*, Toronto, Canada, 2005.

[173] Y. Suzuki and B. van Leer. A Hancock-DG method for hyperbolic-relaxation equations. In 18^{th} *AIAA Computational Fluid Dynamics Conference*, Miami, FL, 2007. (*to appear*).

[174] J. C. Tannehill, D. A. Anderson, and R. H. Pletcher. *Computational Fluid Mechanics and Heat Transfer*. Taylor & Francis, Philadelphia, 2^{nd} edition, 1997.

[175] E. Thiémard. Computing bounds for the star discrepancy. *Computing*, 65:169–186, 2000.

[176] L. N. Trefethen and D. Bau, III. *Numerical Linear Algebra*. SIAM, Philadelphia, 1997.

[177] J. Tsai and L. Lin. A thermal-bubble-actuated micronozzle-diffuser pump. *Journal of Microelectromechanical Systems*, 11(6):665–671, 2002.

[178] J. G. van der Corput. Verteilungsfunktionen I, II. *Koninklijke Nederlandse Akademie van Wetenschappen. Proceedings. Series B.*, 38:813–821, 1058–1066, 1935. (*in Dutch*).

[179] W. G. Vincenti and C. H. Kruger, Jr. *Introduction to Physical Gas Dynamics*. Kreiger, Malabar, FL, 1965.

[180] M. von Smoluchowski. Ueber Wärmeleitung in verdünnten Gasen. *Annalen der Physik*, 300(1):101–130, 1898. (*in German*).

[181] M. von Smoluchowski. Zur kinetischen Theorie der Transpiration und Diffusion verdünnter Gase. *Annalen der Physik*, 338(16):1559–1570, 1910. (*in German*).

[182] W. Wagner. A convergence proof for Bird's direct simulation Monte Carlo method for the Boltzmann equation. *Journal of Statistical Physics*, 66(3/4):1011–1044, 1992.

[183] T. T. Warnock. Computational investigations of low-discrepancy point sets. In S. K. Zaremba, editor, *Applications of Number Theory to Numerical Analysis*, pages 319–343. Academic Press, New York, 1972.

[184] K. Watanabe and H. Komiyama. Micro/macrocavity method applied to the study of the step coverage formation mechanism of SiO_2 films by LPCVD. *Journal of The Electrochemical Society*, 137(4):1222–1227, 1990.

[185] D. Wells. *Prime Numbers: The most mysterious figures in math*. Wiley, New York, 2005.

[186] H. Weyl. Über die Gleichverteilung von Zahlen mod. Eins. *Mathematische Annalen*, 77:313–352, 1916. (*in German*).

[187] F. M. White. *Viscous Fluid Flow*. McGraw-Hill, New York, 2^{nd} edition, 1991.

[188] D. R. Willis. Comparison of kinetic theory analyses of linearized Couette flow. *The Physics of Fluids*, 5(2):127–135, 1962.

[189] S. Wolfram. *The Mathematica Book*. Wolfram Media, 5^{th} edition, 2003.

[190] H. Woźniakowski. Average case complexity of multivariate integration. *Bulletin of the American Mathematical Society*, 24:1985, 1991.

[191] N. Yamanishi, Y. Matsumoto, and K. Shobatake. Multistage gas-surface interaction model for the direct simulation Monte Carlo method. *Physics of Fluids*, 11(11):3540–3552, 1999.

[192] S. K. Zaremba. Applications of multidimensional integration by parts. *Annales Polonici Mathematici*, 21:85–96, 1968.

[193] S. K. Zaremba. Good lattice points in the sense of Hlawka and Monte-Carlo integration. *Monatshefte für Mathematik*, 72:264–269, 1968.

[194] P. Zinterhof. Einige zahlentheoretische Methoden zur numerischen Quadratur und Interpolation. *Österreichische Akademie der Wissenschaften. Mathematisch-Naturwissenschaftliche Klasse. Anzeiger*, 177:51–77, 1969. (*in German*).